From Parameter Estimation to Dispersion of Nonstationary Gauss-Markov Processes

Peida Tian, *Student Member, IEEE,* Victoria Kostina, *Member, IEEE*

***Abstract*—This paper provides a precise error analysis for the maximum likelihood estimate $\hat{a}_{\mathrm{ML}}(u_1^n)$ of the parameter a given samples $u_1^n = (u_1, \ldots, u_n)'$ drawn from a nonstationary Gauss-Markov process $U_i = aU_{i-1} + Z_i,\ i \geq 1$, where $U_0 = 0$, $a > 1$, and Z_i's are independent Gaussian random variables with zero mean and variance σ^2. We show a tight nonasymptotic exponentially decaying bound on the tail probability of the estimation error. Unlike previous works, our bound is tight already for a sample size of the order of hundreds. We apply the new estimation bound to find the dispersion for lossy compression of nonstationary Gauss-Markov sources. We show that the dispersion is given by the same integral formula that we derived previously for the asymptotically stationary Gauss-Markov sources, i.e., $|a| < 1$. New ideas in the nonstationary case include a deeper understanding of the scaling of the maximum eigenvalue of the covariance matrix of the source sequence, and new techniques in the derivation of our estimation error bound.**

***Index Terms*—Parameter estimation, maximum likelihood estimator, unstable processes, finite blocklength analysis, lossy compression, sources with memory, rate-distortion theory, covering in stochastic processes, adaptive control.**

I. Introduction

A. Overview

We consider two related problems that concern a scalar Gauss-Markov process $\{U_i\}_{i=1}^{\infty}$, defined by $U_0 = 0$ and

$$U_i = aU_{i-1} + Z_i, \quad \forall i \geq 1, \tag{1}$$

where Z_i's are independent Gaussian random variables with zero mean and variance σ^2.

The first problem is parameter estimation: given sample u_1^n drawn from the Gauss-Markov source, we seek to design and analyse estimators for the unknown system parameter a. The consistency and asymptotic distribution of the maximum likelihood (ML) estimator have been studied in the literature [2–7]. Our main contribution is a large deviation bound on the estimation error of the ML estimator. Our numerical experiments indicate that our new bound is tighter than previously known results [8–10].

The second problem is the nonasymptotic performance of the optimal lossy compressors of the Gauss-Markov process. An encoder outputs nR bits for each realization u_1^n. Once the decoder receives the nR bits, it produces $\hat{u}_1^n$ as a reproduction of u_1^n. The distortion between u_1^n and $\hat{u}_1^n$ is measured by the mean squared error (MSE). Two commonly used criteria to quantify the distortion of a lossy compression scheme are the average distortion criterion and the excess-distortion probability criterion. The rate-distortion theory, initiated by Shannon [11] and extensively investigated by researchers [12–17], studies the optimal tradeoff between the rate R and the distortion. In the limit of large blocklength n, the minimum rate R required to achieve average distortion d is given by the rate-distortion function. The nonasymptotic version of the rate-distortion problem [18–22] studies the rate-distortion tradeoff for finite blocklength n. Our main contribution is a coding theorem that characterize the gap between the rate-distortion function and the minimum rate R at blocklength n for the nonstationary Gauss-Markov source ($a > 1$), under the excess-distortion probability criterion. We leverage our result on the ML estimator to understand lossy compression as follows. We apply our bound on the estimation error of the ML estimator to construct a typical set of the sequences whose estimated parameter a is close to the true a. We then use the typical set in our achievability proof of the nonasymptotic coding theorem.

Without loss of of generality, we assume that $a \geq 0$ in this paper, since, otherwise, we can consider another random process $\{U_i'\}_{i=1}^{\infty}$ defined by the invertible mapping $U_i' \triangleq (-1)^i U_i$ that satisfies $U_i' = (-a)U_{i-1}' + (-1)^i Z_i$, where $(-1)^i Z_i$'s are also independent zero-mean Gaussian random variables with variance σ^2. We distinguish the following three cases:

- $0 < a < 1$: the asymptotically stationary case;
- $a = 1$: the unit-root case;
- $a > 1$: the nonstationary case.

In this paper, we mostly focus on the nonstationary case.

B. Motivations

Estimation of parameters of stochastic processes from their realizations has many applications. In the statistical analysis of economic time series [2, 23, 24], the Gauss-Markov process $\{U_i\}_{i=1}^{\infty}$ is used to model the varying prices of a certain commodity at time i, and the ML estimate of the unknown coefficient a is then used to predict future prices. [25] and [26, Sec. 5] used the Gauss-Markov process with $a = 1$ to model the stochastic structure of the velocity of money. The Gauss-Markov process, also known as the autoregressive process of order 1 (AR(1)), is a special case of the general autoregressive-moving-average (ARMA) model [27, 28], for which various estimation and prediction procedures have been proposed, e.g. the Box-Jenkins method [28]. The Gauss-Markov process is also a special case of the linear state-space model (e.g. [29, Chap. 5]) that is popular in control theory. One of the problems

P. Tian and V. Kostina are with the Department of Electrical Engineering, California Institute of Technology. (e-mail: {ptian, vkostina}@caltech.edu). This research was supported in part by the National Science Foundation (NSF) under Grant CCF-1751356. A preliminary version [1] of this paper was accepted for publication in the IEEE International Symposium on Information Theory, Paris, France, July 2019.

in control is system identification [30], which is the problem of building mathematical models using measured data from unknown dynamical systems. Parameter estimation is one of the common methods used in system identification where the dynamical system is modeled by a state-space model [30, Chap. 7] with unknown parameters. In modern data-driven control systems, where the goal is to control an unknown nonstationary system given measured data, parameter estimation methods are used as a first step in designing controllers [10] [31, Sec. 1.2]. In speech signal processing, the linear predictive coding algorithm [32] relies on parameter estimation (the ordinary least squares estimate, or, equivalently, the maximum likelihood estimate assuming Gaussian noise) to fit a higher-order Gauss-Markov process, see [32, App. C]. A fine-grained analysis of the ML estimate is instrumental in optimizing the design of all these systems. Our nonasymptotic analysis leading up to a large deviation bound for the ML estimate in our simple setting can provide insights for analyzing more complex random processes, e.g., higher-order autoregressive processes and vector systems.

Understanding finite-blocklength lossy compression of the Gauss-Markov process fits into a continuing effort by many researchers to advance the rate-distortion theory of information sources with memory, see [13–17, 33–43], as well as into a newer push [18–22, 44–49] to understand the fundamental limits of low latency communication. There is a tight connection between lossy compression of the nonstationary Gauss-Markov process and control of an unstable linear system under communication constraints [50, 51]. Namely, the minimum channel capacity needed to achieve a given LQG (linear quadratic Gaussian) cost for the plant [50, Eq. (1)] is lower-bounded by the causal rate-distortion function of the Gauss-Markov process [50, Eq. (9)]. See [51, Th. 1] for more details. Being more restrictive on the coding schemes, the causal rate-distortion function is further lower-bounded by the traditional rate-distortion function. The result in this paper on the rate-distortion tradeoff in the finite blocklength regime provides a lower bound on the minimum communication rate required to ensure that the LQG cost stays below a desired threshold with desired probability at the end of a finite horizon. Finally, the aforementioned linear predictive coding algorithm [32] is connected to lossy compression of autoregressive processes, see a recent historical note by Gray [52, p.2].

C. Notations

For $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, ..., n\}$. We use the standard notations for the asymptotic behaviors $O(\cdot), o(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ and $\omega(\cdot)$. Namely, let $f(n)$ and $g(n)$ be two functions of n, then $f(n) = O(g(n))$ means that there exists a constant $c > 0$ and $n_0 \in \mathbb{N}$ such that $|f(n)| \le M|g(n)|$ for any $n \ge n_0$; $f(n) = o(g(n))$ means $\lim_{n\to\infty} f(n)/g(n) = 0$; $f(n) = \Theta(g(n))$ means there exist positive constants c_1, c_2 and $n_0 \in \mathbb{N}$ such that $c_1 g(n) \le f(n) \le c_2 g(n)$ for any $n \ge n_0$; $f(n) = \Omega(g(n))$ if and only if $g(n) = O(f(n))$; and $f(n) = \omega(g(n))$ if and only if $\lim_{n\to\infty} f(n)/g(n) = +\infty$. For a matrix M, we denote by M' its transpose, by $\|\mathsf{M}\|$ its operator norm (the largest singular value) and by $\mu_1(\mathsf{M}) \le \ldots \le \mu_n(\mathsf{M})$ its eigenvalues listed in nondecreasing order. We use $\mathcal{S}^c$ to denote the complement of a set $\mathcal{S}$. All logarithms and exponentials are base e.

II. Previous Works

A. Parameter Estimation

The maximum likelihood (ML) estimate $\hat{a}_{\mathrm{ML}}(u_1^n)$ of the parameter a given samples $u_1^n = (u_1, \ldots, u_n)'$ drawn from the Gauss-Markov source is given by

$$\hat{a}_{\mathrm{ML}}(u_1^n) = \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\sum_{i=1}^{n-1} u_i^2}. \tag{2}$$

The derivation of (2) is straightforward, e.g. [47, App. F-A]. The problem is to provide performance guarantees of $\hat{a}_{\mathrm{ML}}(u_1^n)$. This simply formulated problem has been widely studied in the literature. Our main contribution in this paper is a nonasymptotic fine-grained large deviations analysis of the estimation error.

The estimate $\hat{a}_{\mathrm{ML}}(u_1^n)$ in (2) has been extensively studied in the statistics [4, 6] and economics [2, 3] communities. Mann and Wald [2] and Rubin [3] showed that the estimation error $\hat{a}_{\mathrm{ML}}(U_1^n) - a$ converges to 0 in probability for any $a \in \mathbb{R}$. Rissanen and Caines [6] later proved that $\hat{a}_{\mathrm{ML}}(U_1^n) - a$ converges to 0 almost surely for $0 < a < 1$. To better understand the finer scaling of the error $\hat{a}_{\mathrm{ML}}(U_1^n) - a$, researchers turned to study the limiting distribution of the normalized estimation error $h(n)(\hat{a}_{\mathrm{ML}}(U_1^n) - a)$ for a careful choice of the standardizing function $h(n)$:

$$h(n) \triangleq \begin{cases} \sqrt{\frac{n}{1-a^2}}, & |a| < 1, \\ \frac{n}{\sqrt{2}}, & |a| = 1, \\ \frac{|a|^n}{a^2-1}, & |a| > 1. \end{cases} \tag{3}$$

With the above choices of $h(n)$, Mann and Wald [2] and White [4] showed that the distribution of the normalized estimation error $h(n)(\hat{a}_{\mathrm{ML}}(U_1^n) - a)$ converges to $\mathcal{N}(0, 1)$ for $|a| < 1$; to the standard Cauchy distribution for $|a| > 1$; and for $|a| = 1$, to the distribution of

$$\frac{B^2(1) - 1}{2\int_0^1 B^2(t)\, dt}, \tag{4}$$

where $\{B(t) : t \in [0, 1]\}$ is a Brownian motion.

Generalizations of the above results in several directions have also been investigated. In [2, Sec. 4], the maximum likelihood estimator for the p-th order stationary autoregressive processes with Z_i's being i.i.d. zero-mean and bounded moments random variables (not necessarily Gaussian) was shown to be weakly consistent, and the scaled estimation errors $\sqrt{n}(\hat{a}_j - a_j)$ for $j = 1, \ldots, p$ were shown to converge in distribution to the Gaussian random variables as n tends to infinity. Anderson [5, Sec. 3] studied the limiting distribution of the maximum likelihood estimator for a nonstationary vector version of the process (1). Chan and Wei [7] studied the performance of the estimation error when a is not a constant but approaches to 1 from below in the order of $1/n$. The problem of estimating the parameter a from a block of outcomes of the Gauss-Markov source (1) is one of the simplest versions in recent studies of machine learning for dynamical systems [10, 53–56]. One objective of

those studies is to obtain tight performance bounds on the least-squares estimates of the system parameters $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ from a single input / output trajectory $\{W_i, Y_i\}_{i=1}^n$ in the following state-space model, e.g. [54, Eq. (1)–(2)]:

$$X_{i+1} = \mathsf{A}X_i + \mathsf{B}W_i + Z_i, \tag{5}$$
$$Y_i = \mathsf{C}X_i + \mathsf{D}W_i + V_i, \tag{6}$$

where X_i, W_i, Z_i, V_i's are random vectors of certain dimensions and the system parameters $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ are matrices of appropriate dimensions. The Gauss-Markov process in (1) can be written as the state-space model by choosing $\mathsf{A} = a$ being a scalar, $\mathsf{B} = \mathsf{D} = 0$, $\mathsf{C} = 1$ and $V_i = 0$. For stable vector systems, that is, $\|\mathsf{A}\| < 1$, Oymak and Ozay [54, Thm. 3.1] showed that the estimation error in spectral norm is $O(1/\sqrt{n})$ with high probability, where n is the number of samples. For the subclass of the regular unstable systems [56, Def. 3], Faradonbeh et al. [56, Thm. 1] proved that the probability of estimation error exceeding a positive threshold in spectral norm decays exponentially in n. For the Gauss-Markov processes considered in the present paper, Simchowitz et al. [53, Thm. B.1] and Sarkar and Rakhlin [55, Prop. 4.1] presented tail bounds on the estimation error of the ML estimate.

Another line of work closely related to this paper is the large deviation principle (LDP) [57, Ch. 1.2] on $\hat{a}_{\mathrm{ML}}(U_1^n) - a$. Given an error threshold $\eta > 0$, define $P^+(n, a, \eta)$ and $P^-(n, a, \eta)$ as follows:

$$P^+(n, a, \eta) \triangleq -\frac{1}{n} \log \mathbb{P}\left[\hat{a}_{\mathrm{ML}}(U_1^n) - a > \eta\right], \tag{7}$$
$$P^-(n, a, \eta) \triangleq -\frac{1}{n} \log \mathbb{P}\left[\hat{a}_{\mathrm{ML}}(U_1^n) - a < -\eta\right]. \tag{8}$$

We also define $P(n, a, \eta)$ as

$$P(n, a, \eta) \triangleq -\frac{1}{n} \log \mathbb{P}\left[|\hat{a}_{\mathrm{ML}}(U_1^n) - a| > \eta\right]. \tag{9}$$

The large deviation theory studies the rate functions, defined as the limits of $P^+(n, a, \eta)$, $P^-(n, a, \eta)$ and $P(n, a, \eta)$, as n goes to infinity. Bercu et al. [8, Prop. 8] found the rate function for the case of $0 < a < 1$. For $a \geq 1$, Worms [9, Thm. 1] proved that the rate functions can be bounded from below implicitly by the optimal value of an optimization problem.

These studies of the limiting distribution and the LDP of the estimation error are both asymptotic. In this paper, we consider the nonasymptotic analysis of the estimation error. Two nonasymptotic lower bounds on $P^+(n, a, \eta)$ and $P^-(n, a, \eta)$ are available in the literature. For any $a \in \mathbb{R}$, Rantzer [10, Th. 4] showed that

$$P^+(n, a, \eta) \ \ (\text{and } P^-(n, a, \eta)) \ \geq \frac{1}{2}\log(1 + \eta^2). \tag{10}$$

Bercu and Touati [58, Cor. 5.2] proved that

$$P^+(n, a, \eta) \ \ (\text{and } P^-(n, a, \eta)) \ \geq \frac{\eta^2}{2(1 + y_\eta)}, \tag{11}$$

where y_η is the unique positive solution to $(1+x)\log(1+x) - x - \eta^2 = 0$ in x. Both bounds (10) and (11) do not capture the dependence on a and n, and are the same for $P^+(n, a, \eta)$ and $P^-(n, a, \eta)$. All the bounds in [10, 53–56] are either optimal only order-wise or involve implicit constants. Our main result on parameter estimation is a tight nonasymptotic lower bound on $P^+(n, a, \eta)$ and $P^-(n, a, \eta)$. For larger a, the lower bound becomes larger, which suggests that unstable systems are easier to estimate than stable ones, an observation consistent with [53]. The proof is inspired by Rantzer [10, Lem. 5], but our result improves Rantzer's result (10) and Bercu and Touati's result (11), see Fig. 1 for a comparison. Most of our results generalize to the case where Z_i's are i.i.d. sub-Gaussian random variables, see Theorem 4 in Section III-D below.

B. Nonasymptotic Rate-distortion Theory

The rate-distortion theory studies the problem of compressing a generic random process $\{X_i\}_{i=1}^\infty$ with minimum distortion. Given a distortion threshold $d > 0$, an excess-distortion probability $\epsilon \in (0, 1)$ and the number of codewords $M \in \mathbb{N}$, an (n, M, d, ϵ) lossy compression code for a random vector X_1^n consists of an encoder $\mathsf{f}_n \colon \mathbb{R}^n \to [M]$, and a decoder $\mathsf{g}_n \colon [M] \to \mathbb{R}^n$, such that $\mathbb{P}\left[\mathsf{d}\left(X_1^n, \mathsf{g}_n\left(\mathsf{f}_n(X_1^n)\right)\right) > d\right] \leq \epsilon$, where $\mathsf{d}(\cdot, \cdot)$ is the distortion measure. This paper considers the mean squared error (MSE) distortion: $\forall\ x_1^n,\ y_1^n \in \mathbb{R}^n$,

$$\mathsf{d}(x_1^n, y_1^n) \triangleq \frac{1}{n}\sum_{i=1}^n (x_i - y_i)^2. \tag{12}$$

The minimum achievable code size and source coding rate are defined respectively by

$$M^\star(n, d, \epsilon) \triangleq \min\left\{M \in \mathbb{N} \colon \exists\ (n, M, d, \epsilon)\ \text{code}\right\}, \tag{13}$$
$$R(n, d, \epsilon) \triangleq \frac{1}{n}\log M^\star(n, d, \epsilon). \tag{14}$$

In this paper, we approximate the nonasymptotic coding rate $R(n, d, \epsilon)$ for the nonstationary Gauss-Markov source.

Another related and widely studied setting is compression under the average distortion criterion. Given a distortion threshold $d > 0$ and the number of codewords $M \in \mathbb{N}$, an (n, M, d) lossy compression code for a random vector X_1^n consists of an encoder $\mathsf{f}_n \colon \mathbb{R}^n \to [M]$, and a decoder $\mathsf{g}_n \colon [M] \to \mathbb{R}^n$, such that $\mathbb{E}\left[\mathsf{d}\left(X_1^n, \mathsf{g}_n\left(\mathsf{f}_n(X_1^n)\right)\right)\right] \leq d$. Similarly, one can define $M^\star(n, d)$ and $R(n, d)$ as the minimum achievable code size and source coding rate, respectively, under the average distortion criterion. The traditional rate-distortion theory [11, 12, 14, 15, 34, 59] showed that the limit of the operational source coding rate $R(n, d)$ as n tends to infinity equals the informational rate-distortion function for a wide class of sources. For discrete memoryless sources, Zhang, Yang and Wei in [19] showed that $R(n, d)$ approaches the rate-distortion function in the order $\log n/2n + o(\log n/n)$. For abstract alphabet memoryless sources, Yang and Zhang in [20, Th. 2] showed a similar convergence rate.

Under the excess-distortion probability criterion, one can also study the nonasymptotic behavior of the minimum achievable excess-distortion probability $\epsilon^\star(n, d, M)$:

$$\epsilon^\star(n, d, M) \triangleq \inf\left\{\epsilon > 0 \colon \exists\ (n, M, d, \epsilon)\ \text{code}\right\}. \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

Marton's excess distortion exponent [18, Th. 1, Eq. (2)-(3), (20)] showed that for discrete memoryless sources P_X, it holds that

$$-\frac{1}{n}\log\epsilon^\star(n,d,M)=\min_{P_{\hat{X}}}\ D(P_{\hat{X}}||P_X)+O\left(\frac{\log n}{n}\right), \tag{16}$$

where the minimization is over all probability distributions $P_{\hat{X}}$ such that $\mathbb{R}_{\hat{X}}(d)\geq\frac{\log M}{n}$, where M is such that $\frac{\log M}{n}$ is a constant, $\mathbb{R}_{\hat{X}}(d)$ denotes the rate-distortion function of a discrete memoryless source with single-letter distribution $P_{\hat{X}}$, and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence. As pointed out by [21, p.2], for fixed $d>0$ and $\epsilon\in(0,1)$, even the asymptotic behavior of $R(n,d,\epsilon)$ is unanswered by Marton's bound in (16). Ingber and Kochman [21] (for finite-alphabet and Gaussian sources) and Kostina and Verdú [22] (for abstract sources) showed that the minimum achievable source coding rate $R(n,d,\epsilon)$ satisfies a Gaussian approximation:

$$R(n,d,\epsilon)\approx\mathbb{R}_X(d)+Q^{-1}(\epsilon)\sqrt{\frac{\mathbb{V}(d)}{n}}, \tag{17}$$

where $\mathbb{V}(d)$ is the dispersion of the source (defined as the variance of the tilted information random variable, details later) and Q^{-1} denotes the inverse q-function. In this paper, by extending our previous analysis [47, Th. 1] of the stationary Gauss-Markov source to the nonstationary one, we establish a Gaussian approximation in the form of (17) for the nonstationary Gauss-Markov sources. One of the key ideas behind this extension is to construct a typical set using the ML estimate of a, and to use our estimation error bound to probabilistically characterize that set.

III. Parameter Estimation

A. Nonasymptotic Lower Bounds

We first present our nonasymptotic bounds on $P^+(n,a,\eta)$ and $P^-(n,a,\eta)$, defined in (7) and (8) above, respectively. We define two sequences $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ and $\{\beta_\ell\}_{\ell\in\mathbb{N}}$ as follows. Let $\sigma^2>0$ and $a>1$ be fixed constants. For $\eta>0$ and a parameter $s>0$, let α_ℓ be the following sequence

$$\alpha_1\triangleq\frac{\sigma^2s^2-2\eta s}{2}, \tag{18}$$

$$\alpha_\ell=\frac{[a^2+2\sigma^2s(a+\eta)]\alpha_{\ell-1}+\alpha_1}{1-2\sigma^2\alpha_{\ell-1}},\quad\forall\ell\geq2. \tag{19}$$

Similarly, let β_ℓ be the following sequence

$$\beta_1\triangleq\frac{\sigma^2s^2-2\eta s}{2}, \tag{20}$$

$$\beta_\ell=\frac{[a^2+2\sigma^2s(-a+\eta)]\beta_{\ell-1}+\beta_1}{1-2\sigma^2\beta_{\ell-1}},\quad\forall\ell\geq2. \tag{21}$$

Note the subtle difference between (19) and (21): there is a negative sign in the numerator in (21). Both sequences depend on η and s. We derive closed-form expressions and analyze the convergence properties of α_ℓ and β_ℓ in Appendices A-B and A-C below. For $\eta>0$ and $n\in\mathbb{N}$, we define the following sets

$$\mathcal{S}_n^+\triangleq\left\{s\in\mathbb{R}\colon s>0,\ \alpha_\ell<\frac{1}{2\sigma^2},\ \forall\ell\in[n]\right\}, \tag{22}$$

$$\mathcal{S}_n^-\triangleq\left\{s\in\mathbb{R}\colon s>0,\ \beta_\ell<\frac{1}{2\sigma^2},\ \forall\ell\in[n]\right\}. \tag{23}$$

Theorem 1. *For any constant $\eta>0$, the estimator (2) satisfies for any $n\geq2$,*

$$P^+(n,a,\eta)\geq\sup_{s\in\mathcal{S}_n^+}\ \frac{1}{2n}\sum_{\ell=1}^{n-1}\log\left(1-2\sigma^2\alpha_\ell\right), \tag{24}$$

$$P^-(n,a,\eta)\geq\sup_{s\in\mathcal{S}_n^-}\ \frac{1}{2n}\sum_{\ell=1}^{n-1}\log\left(1-2\sigma^2\beta_\ell\right), \tag{25}$$

where α_ℓ and β_ℓ are defined in (19) and (21), respectively, and $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$ are defined in (22) and (23), respectively.

Theorem 1 is a useful result for numerically computing lower bounds on $P^+(n,a,\eta)$ and $P^-(n,a,\eta)$. In Fig. 1, we plot our lower bounds in Theorem 1, previous results in (10) by Rantzer and (11) by Bercu and Touati, and a simulation result. As one can see, our bound in Theorem 1 is much tighter than previous results.

The proof of Theorem 1, presented in Appendix A-A below, is a detailed analysis of the Chernoff bound using the tower property of conditional expectations. The proof is motivated by [10, Lem. 5], but our analysis is more accurate and the result is significantly tighter, see Fig. 1 and Fig. 3 for comparisons. One recovers Rantzer's lower bound (10) by setting $s=\eta/\sigma^2$ and bounding α_ℓ as $\alpha_\ell\leq\alpha_1$ (due to the monotonicity of α_ℓ shown in Appendix A-B below) in Theorem 1. We explicitly state where we diverge from [10, Lem. 5] in the proof in Appendix A-A below.

Remark 1. In view of the Gärtner-Ellis theorem [57, Th. 2.3.6], we conjecture that the bounds (24) and (25) can be reversed in the limit of large n:

$$\limsup_{n\to\infty}P^+(n,a,\eta)\leq\limsup_{n\to\infty}\sup_{s\in\mathcal{S}_n^+}\ \frac{1}{2n}\sum_{\ell=1}^{n-1}\log\left(1-2\sigma^2\alpha_\ell\right), \tag{26}$$

and similarly for (25).

B. Asymptotic Lower Bounds

We next present our bounds on the error exponents, that is, the limits of $P^+(n,a,\eta)$, $P^-(n,a,\eta)$ and $P(n,a,\eta)$ as n tends to infinity. To take limits using (24) and (25), we need to understand the two sequences of sets $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$. Define the limits of the sets as

$$\mathcal{S}_\infty^+\triangleq\bigcap_{n\geq1}\mathcal{S}_n^+, \tag{27}$$

$$\mathcal{S}_\infty^-\triangleq\bigcap_{n\geq1}\mathcal{S}_n^-. \tag{28}$$

We have the following properties.

Lemma 1. *Fix any constant $\eta>0$.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

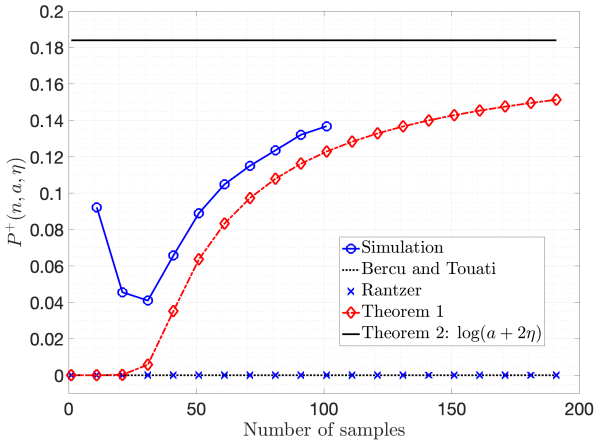

Fig. 1: Numerical simulations and lower bounds on $P^+(n,a,\eta)$: We choose $a = 1.2$ and $\eta = 10^{-3}$. For each n, we generate $N = 10^6$ independent samples u_1^n from the Gauss-Markov process (1). We approximate $P^+(n,a,\eta)$ by $-\frac{1}{n}\log\left(\frac{1}{N}\#\{\text{samples } u_1^n \text{ with } \hat{a}_{\text{ML}}(u_1^n) - a > \eta\}\right)$. We plot lower bounds on $P^+(n,a,\eta)$ by Rantzer (10), Bercu and Touati in (11), our nonasymptotic bound in (24) and the asymptotic bound in Theorem 2 in Section III-B below.

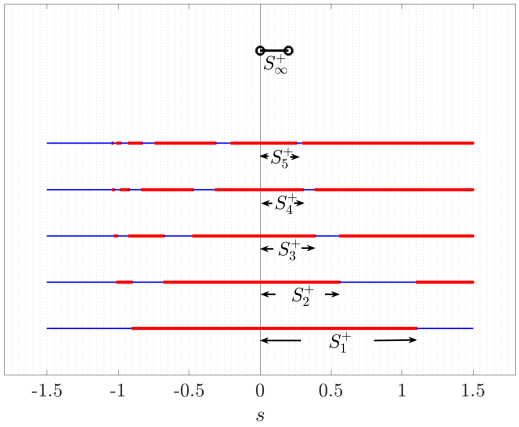

Fig. 2: Numerical computation of the sets $\mathcal{S}_n^+$ for $a = 1.2$ and $\eta = 0.1$. Each horizontal line corresponds to $n = 1, ..., 5$ in the bottom-up order. Within each horizontal line, the red thick parts denote the ranges of s for which $\alpha_n < \frac{1}{2\sigma^2}$, and the blue thin region is where $\alpha_n \geq \frac{1}{2\sigma^2}$. The plot for $\mathcal{S}_n^-$ is similar.

- *(Monotone decreasing sets) For any $n \geq 1$, we have*

$$\mathcal{S}_{n+1}^+ \subseteq \mathcal{S}_n^+, \quad \mathcal{S}_{n+1}^- \subseteq \mathcal{S}_n^-. \tag{29}$$

- *(Limits of the sets) It holds that*

$$\mathcal{S}_\infty^+ = \left(0, \frac{2\eta}{\sigma^2}\right], \tag{30}$$

$$\mathcal{S}_\infty^- \supsetneq \left(0, \frac{2\eta}{\sigma^2}\right]. \tag{31}$$

The proof of Lemma 1 is presented in Appendix A-D below. The exact characterization of $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$ for each n using η is involved. One can see from the definitions (22) and (23) that

$$\mathcal{S}_1^+ = \mathcal{S}_1^- = \left\{s \in \mathbb{R} \colon 0 < s < \frac{\eta + \sqrt{1+\eta^2}}{\sigma^2}\right\}. \tag{32}$$

To obtain the set $\mathcal{S}_{n+1}^+$ from $\mathcal{S}_n^+$, we need to solve $\alpha_{n+1} < \frac{1}{2\sigma^2}$, which is equivalent to solving an additional inequality involving a polynomial of degree $n+2$ in s (using the closed-form expression for α_{n+1} in (128) in Appendix A-B below). Fig. 2 presents a plot of $\mathcal{S}_n^+$ for $n = 1, ..., 5$. Despite the complexity of the sets $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$, Lemma 1 shows their monotonicity property and limits.

Combining Theorem 1 and Lemma 1, we obtain the following lower bounds on the error exponents. The proof is given in Appendix A-E below.

Theorem 2. *Fix any constant $\eta > 0$. For the ML estimator (2), the following three inequalities hold:*

$$\liminf_{n\to\infty} \; P^+(n,a,\eta) \geq I^+(a,\eta) \triangleq \log(a+2\eta), \tag{33}$$

$$\liminf_{n\to\infty} \; P^-(n,a,\eta) \geq I^-(a,\eta), \tag{34}$$

$$\liminf_{n\to\infty} \; P(n,a,\eta) \geq I^-(a,\eta), \tag{35}$$

where

$$I^-(a,\eta) \triangleq \begin{cases} \log a, & 0 < \eta \leq \eta_1, \\ \frac{1}{2}\log\frac{2a\eta - (a^2-1)}{1-(\eta-a)^2}, & \eta_1 < \eta < \eta_2, \\ \log(2\eta - a), & \eta \geq \eta_2, \end{cases} \tag{36}$$

with the thresholds η_1 and η_2 given by

$$\eta_1 \triangleq \frac{a^2-1}{a}, \tag{37}$$

$$\eta_2 \triangleq \frac{3a + \sqrt{a^2+8}}{4}. \tag{38}$$

Remark 2. The results in (30)-(31) and (33)-(34) indicate the asymmetry between $P^+(n,a,\eta)$ and $P^-(n,a,\eta)$: the set $\mathcal{S}_\infty^-$ has a larger range than $\mathcal{S}_\infty^+$, and $I^+(a,\eta) > I^-(a,\eta)$, which suggests that the maximum likelihood estimator $\hat{a}_{\text{ML}}(U_1^n)$ is more likely to underestimate a than to overestimate it.

Fig. 3 presents a comparison of (35), Rantzer's bound (10) and Bercu and Touati (11). Our bound (35) is tighter than both of them for any $\eta > 0$.

C. Decreasing Error Thresholds

When the number of samples n increases, it is natural to also have smaller error thresholds η. In this section, we consider the regime where the error threshold $\eta = \eta_n > 0$ is a sequence decreasing to 0. In this setting, Theorem 1 still holds and the proof stays the same, except that we replace α_ℓ and β_ℓ, by the length-n sequences $\alpha_{n,\ell}$ and $\beta_{n,\ell}$ for $\ell = 1, \ldots, n$, respectively, where $\alpha_{n,\ell}$ and $\beta_{n,\ell}$ now depend on η_n instead of a constant η:

$$\alpha_{n,1} \triangleq \frac{\sigma^2 s^2 - 2\eta_n s}{2}, \tag{39}$$

$$\alpha_{n,\ell} = \frac{[a^2 + 2\sigma^2 s(a+\eta_n)]\alpha_{n,\ell-1} + \alpha_{n,1}}{1 - 2\sigma^2\alpha_{n,\ell-1}}, \quad \forall \ell = 2, \ldots, n. \tag{40}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

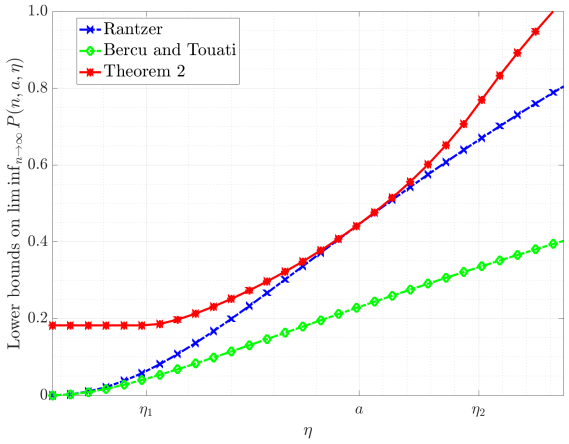

Fig. 3: Comparisons lower bounds on $\liminf_{n\to\infty} P(n,a,\eta)$: For $a = 1.2$, we plot the three lower bounds in Rantzer (10), Bercu and Touati (11) and our (35) in Theorem 2.

The sequence $\beta_{n,\ell}$ is defined in a similar way. For Theorem 2 to remain valid, we require η_n no smaller than $1/\sqrt{n}$ to ensure that the right sides of (24)-(25) still converge to the right sides of (33)-(34), respectively. Let η_n be a positive sequence such that

$$\eta_n = \omega\left(\frac{1}{\sqrt{n}}\right). \tag{41}$$

Theorem 3. *For any $\sigma^2 > 0$ and $a > 1$, let $\eta_n > 0$ be a positive sequence satisfying (41). Then, Theorem 1 holds with α_ℓ replaced by $\alpha_{n,\ell}$, and β_ℓ by $\beta_{n,\ell}$, and Theorem 2 holds with (33) and (34) replaced, respectively, by*

$$\liminf_{n\to\infty} P^+(n,a,\eta_n) \geq \log a, \tag{42}$$

$$\liminf_{n\to\infty} P^-(n,a,\eta_n) \geq \log a. \tag{43}$$

The proof of Theorem 3 is presented in Appendix A-F below. Theorem 3 is a quite strong result as it states that even if the error threshold is a sequence decreasing to zero, as long as (41) is satisfied, the probability of estimation error exceeding such decreasing thresholds is still exponentially small, with exponent being at least $\log a$.

Corollary 1. *For any $\sigma^2 > 0$ and any $a > 1$, there exists a constant $c \geq \frac{1}{2}\log(a)$ such that for all n large enough,*

$$\mathbb{P}\left[|\hat{a}_{ML}(U_1^n) - a| \geq \sqrt{\frac{\log\log n}{n}}\right] \leq 2e^{-cn}. \tag{44}$$

Corollary 1 is used in Section IV-E below to derive the dispersion of nonstationary Gauss-Markov sources. The proof of Corollary 1 is by applying Theorem 3 with η_n chosen as

$$\eta_n = \sqrt{\frac{\log\log n}{n}}. \tag{45}$$

D. Generalization to sub-Gaussian Z_i's

In this section, we generalize the above results to the case where Z_i's in (1) are zero-mean σ-sub-Gaussian random variables. This general result is of independent interest and will not be used in the rest of the paper.

Definition 1 (sub-Gaussian random variable, e.g. [60, Def. 2.7])**.** *Fix $\sigma > 0$. A random variable $Z \in \mathbb{R}$ with mean μ is said to be σ-sub-Gaussian with variance proxy σ^2 if its moment-generating function (MGF) satisfies*

$$\mathbb{E}[e^{s(Z-\mu)}] \leq e^{\frac{\sigma^2 s^2}{2}}, \tag{46}$$

for all $s \in \mathbb{R}$.

One important property of σ-sub-Gaussian random variables is the following well-known bound on the MGF of quadratic functions of σ-sub-Gaussian random variables.

Lemma 2 ([10, Prop. 2])**.** *Let Z be a σ-sub-Gaussian random variable with mean μ. Then*

$$\mathbb{E}\left[\exp(sZ^2)\right] \leq \frac{1}{\sqrt{1-2\sigma^2 s}} \exp\left(\frac{s\mu^2}{1-2\sigma^2 s}\right) \tag{47}$$

for any $s < \frac{1}{2\sigma^2}$.

Equality holds in (46) and (47) when Z is Gaussian. In particular, the right side of (47) is the MGF of the noncentral χ^2-distributed random variable Z^2.

Theorem 4 (Generalization to sub-Gaussian case)**.** *Theorems 1–3 and Lemma 1 remain valid for the estimator (2) when Z_i's in (1) are i.i.d. zero-mean σ-sub-Gaussian random variables.*

The generalizations of Theorems 1–3 and Lemma 1 from Gaussian to sub-Gaussian Z_i's only require minor changes in the corresponding proofs. See Appendix A-G for the details.

IV. The Dispersion of a Nonstationary Gauss-Markov Source

A. Rate-distortion functions

For a generic random process $\{X_i\}_{i=1}^\infty$, the n-th order rate-distortion function $\mathbb{R}_{X_1^n}(d)$ is defined as

$$\mathbb{R}_{X_1^n}(d) \triangleq \inf_{\substack{P_{Y_1^n|X_1^n}:\\ \mathbb{E}[\mathsf{d}(X_1^n,Y_1^n)]\leq d}} \frac{1}{n} I(X_1^n; Y_1^n), \tag{48}$$

where $X_1^n \triangleq (X_1,\ldots,X_n)'$ is the n-dimensional random vector determined by the random process, $I(X_1^n;Y_1^n)$ is the mutual information between X_1^n and Y_1^n, d is a given distortion threshold, and $\mathsf{d}(\cdot,\cdot)$ is the distortion measure defined in (12) in Sec. II-B above. The rate-distortion function $\mathbb{R}_X(d)$ is defined as

$$\mathbb{R}_X(d) \triangleq \limsup_{n\to\infty} \mathbb{R}_{X_1^n}(d). \tag{49}$$

For a wide class of sources, the rate-distortion function $\mathbb{R}_X(d)$ has been shown to be equal to the minimum achievable source coding rate under the average distortion criterion, in the limit of $n \to \infty$, see [11] for discrete memoryless sources and [12] for general ergodic sources. In particular, Gray's coding theorem [15, Th. 2] for the Gaussian autoregressive processes directly implies that for the Gauss-Markov source $\{U_i\}_{i=1}^\infty$ in (1) for any $a \in \mathbb{R}$, its rate-distortion function $\mathbb{R}_U(d)$

equals the minimum achievable source coding rate under the average distortion criterion as n tends to infinity. The n-th order rate-distortion function $\mathbb{R}_{U_1^n}(d)$ of the Gauss-Markov source is given by the n-th order reverse waterfilling, e.g. [15, Eq. (22)]:

$$\mathbb{R}_{U_1^n}(d) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\max\left(\mu_{n,i},\ \frac{\sigma^2}{\theta_n}\right), \tag{50}$$

$$d = \frac{1}{n}\sum_{i=1}^{n}\min\left(\theta_n, \frac{\sigma^2}{\mu_{n,i}}\right), \tag{51}$$

where $\theta_n > 0$ is the n-th order water level, and $\mu_{n,i}$'s for $i \in [n]$ (sorted in nondecreasing order) are the eigenvalues of the $n \times n$ matrix $\mathsf{F}'\mathsf{F}$ with F being an $n \times n$ lower triangular matrix defined as

$$(\mathsf{F})_{ij} \triangleq \begin{cases} 1, & i = j, \\ -a, & i = j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

One can check that $\sigma^2(\mathsf{F}'\mathsf{F})^{-1}$ is the covariance matrix of U_1^n. The way that we use (50)-(51) is to first solve the n-th order water level θ_n using (51) for a given distortion threshold d, then plugging that water level into (50) to obtain $\mathbb{R}_{U_1^n}(d)$. The rate-distortion function $\mathbb{R}_U(d)$ of the Gauss-Markov source is given by the limiting reverse waterfilling:

$$\mathbb{R}_U(d) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log\max\left(g(w),\ \frac{\sigma^2}{\theta}\right)\,dw, \tag{53}$$

$$d = \frac{1}{2\pi}\int_{-\pi}^{\pi}\min\left(\theta,\ \frac{\sigma^2}{g(w)}\right)\,dw, \tag{54}$$

where $\theta > 0$ is the limiting water level and $g(w)$ is a function from $[-\pi, \pi]$ to $\mathbb{R}$ given by

$$g(w) \triangleq 1 + a^2 - 2a\cos(w). \tag{55}$$

Moreover, it is well-known [11] that the rate-distortion function of the Gaussian memoryless source $\{Z_i\}_{i=1}^{\infty}$ (the special case when a is set to 0 in the Gauss-Markov model) is

$$\mathbb{R}_Z(d) = \max\left(0,\ \frac{1}{2}\log\frac{\sigma^2}{d}\right). \tag{56}$$

See Fig. 4 for a plot of $\mathbb{R}_U(d)$ and $\mathbb{R}_Z(d)$.

B. Operational Dispersion

To characterize the convergence rate of the minimum achievable source coding rate $R(n, d, \epsilon)$ (defined in (14) in Section II-B above) to the rate-distortion function, we define the operational dispersion $V_U(d)$ for the Gauss-Markov source as

$$V_U(d) \triangleq \lim_{\epsilon\to 0}\limsup_{n\to\infty} n\left(\frac{R(n,d,\epsilon) - \mathbb{R}_U(d)}{Q^{-1}(\epsilon)}\right)^2, \tag{57}$$

where Q^{-1} denotes the inverse Q-function. The main result in the second part of this paper gives $V_U(d)$ for the nonstationary Gauss-Markov source.

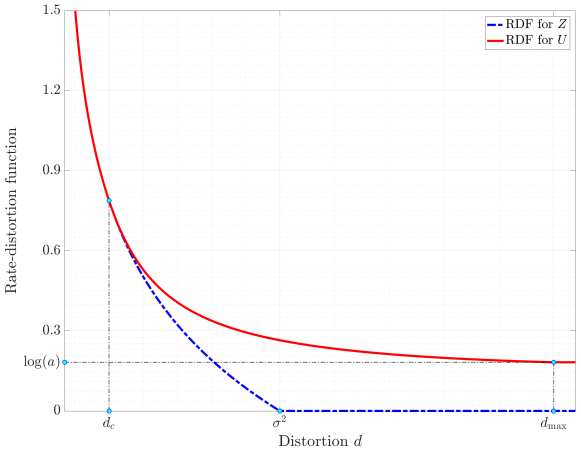

Fig. 4: Rate-distortion functions: $\mathbb{R}_U(d)$ in (53) with $a = 1.2$, and $\mathbb{R}_Z(d)$ in (56).

C. Informational Dispersion

The d-tilted information [22, Def. 6] is the key random variable in our nonasymptotic analysis of $R(n, d, \epsilon)$. Under different names, the d-tilted information has also been studied by Blahut [61, Th. 4] and Kontoyiannis [36, Sec. III-A]. Using the definition in [22, Def. 6], the d-tilted information $\jmath_{U_1^n}(u_1^n, d)$ in u_1^n is

$$\jmath_{U_1^n}(u_1^n, d) \triangleq -\lambda_n^\star d - \log\mathbb{E}\exp\left(-\lambda_n^\star \mathsf{d}(u_1^n, V_1^{\star n})\right), \tag{58}$$

where $\lambda_n^\star$ is the negative slope of $\mathbb{R}_{U_1^n}(d)$ at the distortion level d and $V_1^{\star n}$ is the random variable that achieves the infimum in (48) for U_1^n. In [47, Lem. 7, Eq. (228)], by a decorrelation argument, we obtained the following expression for the d-tilted information for the Gauss-Markov source: for any $a \in \mathbb{R}$ and any $n \in \mathbb{N}$,

$$\jmath_{U_1^n}(u_1^n, d) = \sum_{i=1}^{n}\frac{\min(\theta_n, \sigma_{n,i}^2)}{2\theta_n}\left(\frac{x_i^2}{\sigma_{n,i}^2} - 1\right) + \frac{1}{2}\sum_{i=1}^{n}\log\frac{\max(\theta_n, \sigma_{n,i}^2)}{\theta_n}, \tag{59}$$

where $\theta_n > 0$ is given by (51), $x_1^n \triangleq \mathsf{S}'u_1^n$ with S being an $n \times n$ orthonormal matrix that diagonalizes $(\mathsf{F}'\mathsf{F})^{-1}$, and

$$\sigma_{n,i}^2 \triangleq \frac{\sigma^2}{\mu_{n,i}} \tag{60}$$

with $\mu_{n,i}$'s being the eigenvalues of the $n \times n$ matrix $\mathsf{F}'\mathsf{F}$. We refer to the random variable X_1^n, defined by

$$X_1^n \triangleq \mathsf{S}'U_1^n, \tag{61}$$

as the decorrelation of U_1^n. Note that the decorrelation X_1^n has independent coordinates and

$$X_i \sim \mathcal{N}(0, \sigma_{n,i}^2). \tag{62}$$

Using (50)-(51) and (62), one can show [47, Eq. (55) and (228)] that the d-tilted information $\jmath_{U_1^n}(u_1^n, d)$ in u_1^n for the Gauss-Markov source satisfies $\jmath_{U_1^n}(u_1^n, d) = \jmath_{X_1^n}(x_1^n, d)$. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

minimum achievable source coding rates (defined in (14)) for lossy compression of U_1^n and X_1^n are equal, as are their rate-distortion functions: $\mathbb{R}_{U_1^n}(d) = \mathbb{R}_{X_1^n}(d)$, see [47, Sec. III.A] for the detail. It is known [22, Property 1] that the d-tilted information $\jmath_{U_1^n}(u_1^n, d)$ satisfies (by the Karush-Kuhn-Tucker conditions for the optimization problem (48), roughly speaking)

$$\mathbb{E}\left[\jmath_{U_1^n}(U_1^n, d)\right] = \mathbb{R}_{U_1^n}(d). \tag{63}$$

The informational dispersion $\mathbb{V}_U(d)$ is then defined as the limit of the variance of the d-tilted information normalized by n:

$$\mathbb{V}_U(d) \triangleq \limsup_{n\to\infty} \frac{1}{n}\mathrm{Var}\left[\jmath_{U_1^n}(U_1^n, d)\right]. \tag{64}$$

By decorrelating the Gauss-Markov source U_1^n and analyzing the limiting behavior of the eigenvalues of the covariance matrix of U_1^n, we obtain the following reverse waterfilling representation for the informational dispersion. The proof is given in Appendix B-A below.

Lemma 3. *The informational dispersion of the nonstationary Gauss-Markov source is given by*

$$\mathbb{V}_U(d) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \min\left[1, \left(\frac{\sigma^2}{\theta g(w)}\right)^2\right] dw, \tag{65}$$

where $\theta > 0$ is given in (54), and g is in (55).

Notice that the informational dispersion in the nonstationary case is given by the same expression as in the stationary case [47, Eq. (57)]. It is known, e.g. [22, Eq. (94)] and [21, Sec. IV], that the informational dispersion for the Gaussian memoryless source $\{Z_i\}_{i=1}^{\infty}$ is

$$\mathbb{V}_Z(d) = \frac{1}{2}, \quad \forall d \in (0, \sigma^2). \tag{66}$$

See Fig. 5 for a plot of $\mathbb{V}_U(d)$ and $\mathbb{V}_Z(d)$.

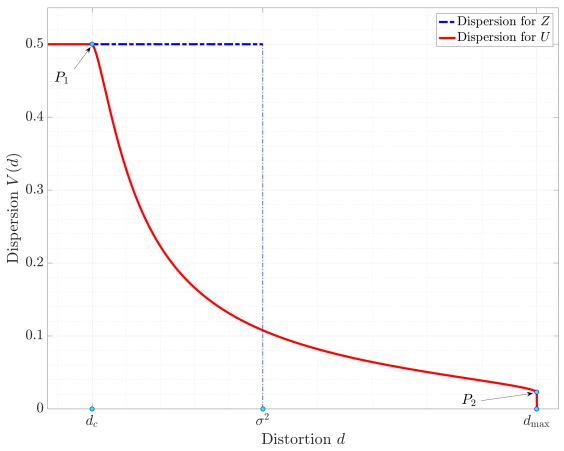

Fig. 5: Dispersions :$\mathbb{V}_U(d)$ in (65) with $a = 1.2$, and $\mathbb{V}_Z(d)$ in (66).

D. A Few Remarks

In view of (54), there are two special water levels $\theta_{\min}$ and $\theta_{\max}$, defined as follows:

$$\theta_{\min} \triangleq \min_{w\in[-\pi,\pi]} \frac{\sigma^2}{g(w)} = \frac{\sigma^2}{(a+1)^2} \tag{67}$$

and

$$\theta_{\max} \triangleq \max_{w\in[-\pi,\pi]} \frac{\sigma^2}{g(w)} = \frac{\sigma^2}{(a-1)^2}. \tag{68}$$

The critical distortion d_c is defined as the distortion corresponding to the water level $\theta_{\min}$, and by (54) we have

$$d_c = \theta_{\min} = \frac{\sigma^2}{(a+1)^2}. \tag{69}$$

The maximum distortion $d_{\max}$ is defined as the distortion corresponding to the water level $\theta_{\max}$. By (54), we have

$$d_{\max} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sigma^2}{g(w)}\, dw. \tag{70}$$

Using similar techniques as in [47, Eq. (169)–(172)], one can compute the integral in (70) as

$$d_{\max} = \frac{\sigma^2}{a^2-1}. \tag{71}$$

In this paper, we always consider a fixed distortion threshold d such that $0 < d < d_{\max}$.

Remark 3. Gray [15, Eq. (24)] showed the following relation between the rate-distortion function $\mathbb{R}_U(d)$ of the Gauss-Markov source and $\mathbb{R}_Z(d)$ of the Gaussian memoryless source:

$$\begin{cases} \mathbb{R}_U(d) = \mathbb{R}_Z(d), & d \in (0, d_c], \\ \mathbb{R}_U(d) > \mathbb{R}_Z(d), & d \in (d_c, d_{\max}). \end{cases} \tag{72}$$

Using Lemma 3 above, one can easily show (in the same way as [47, Cor. 1]) that their dispersions are also comparable:

$$\begin{cases} \mathbb{V}_U(d) = \mathbb{V}_Z(d), & d \in (0, d_c], \\ \mathbb{V}_U(d) < \mathbb{V}_Z(d), & d \in (d_c, \sigma^2). \end{cases} \tag{73}$$

The results in (72)-(73) imply that for low distortions $d \in (0, d_c)$, the minimum achievable source coding rate in compressing the Gauss-Markov source and the Gaussian memoryless source are the same up to second-order terms, a phenomenon we observed in the stationary case as well [47, Cor. 1]. See Fig. 4 and Fig. 5 for a visualization of (72) and (73), respectively.

Remark 4. For the function $\mathbb{R}_U(d)$, we show that

$$\mathbb{R}_U(d_{\max}) = \log a. \tag{74}$$

This result has an interesting connection to the problem of control under communication constraints [62] [63, Th. 1] [64, Prop. 3.1], where it was shown that the minimum rate to asymptotically stabilize a linear, discrete-time, scalar system is also $\log a$, suggesting that stability is unattained with any rate lower than $\log a$ even if an infinite lookahead is allowed. The derivation of (74) is presented in Appendix B-C below.

Remark 5. Let P_1 and P_2 be the two special points on the curve $\mathbb{V}_U(d)$ at distortions d_c and $d_{\max}$, respectively. Then, the coordinates of P_1 and P_2 are given by

$$P_1 = (d_c, 1/2), \quad P_2 = \left(d_{\max}, \frac{(1+a^2)(a-1)}{2(a+1)^3}\right). \tag{75}$$

The derivation for P_2 is the same as that in the stationary case [47, Eq. (61)] except that we need to compute the residue at $1/a$ instead of at a since we now have $a > 1$, see [47, App. B-A] for details.

E. Second-order Coding Theorem

Our main result establishes the equality between the operational dispersion and the informational dispersion.

Theorem 5 (Gaussian approximation)**.** *For the Gauss-Markov source (1) with $a > 1$. For any fixed excess-distortion probability $\epsilon \in (0,1)$ and distortion threshold $d \in (0, d_{\max})$, we have*

$$V_U(d) = \mathbb{V}_U(d). \tag{76}$$

Specifically, we have the following converse and achievability.

Theorem 6 (Converse)**.** *For the Gauss-Markov source with $a > 1$, for any fixed excess-distortion probability $\epsilon \in (0,1)$ and distortion threshold d, the minimum achievable source coding rate $R(n,d,\epsilon)$ satisfies*

$$R(n,d,\epsilon) \geq \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}} Q^{-1}(\epsilon) - \frac{\log n}{2n} + O\left(\frac{1}{n}\right), \tag{77}$$

where Q^{-1} denotes the inverse Q-function, $\mathbb{R}_U(d)$ is the rate-distortion function given in (53), and $\mathbb{V}_U(d)$ is the informational dispersion given by Lemma 3 above.

The converse proof is similar to that in the asymptotically stationary case in [47, Th. 7]. See Appendix D for the details.

Theorem 7 (Achievability)**.** *In the setting the Theorem 6, the minimum achievable source coding rate $R(n,d,\epsilon)$ satisfies*

$$R(n,d,\epsilon) \leq \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}} Q^{-1}(\epsilon) + O\left(\frac{1}{\sqrt{n}\log n}\right). \tag{78}$$

It is straightforward that (76) follows from Theorems 6 and 7. Central to the achievability proof of Theorem 7 is the random coding bound. Specifically, direct application of [22, Cor. 11] implies that there exists an (n, M, d, ϵ) code such that

$$\epsilon \leq \inf_{P_{V_1^n}} \mathbb{E}\left[\exp\left(-M \cdot P_{V_1^n}(\mathcal{B}(U_1^n, d))\right)\right], \tag{79}$$

where the infimization is over all random variables defined on $\mathbb{R}^n$ and $\mathcal{B}(u_1^n, d))$ denotes the distortion d-ball around u_1^n:

$$\mathcal{B}(u_1^n, d)) \triangleq \{z_1^n \in \mathbb{R}^n \colon \mathsf{d}(u_1^n, z_1^n) \leq d\}. \tag{80}$$

To obtain the achievability in (78) from (79), we need to bound from below the probability of the distortion d-ball in terms of the informational dispersion. This connection is made via the following second-order refinement of the "lossy AEP" (asymptotic equipartition property) for the nonstationary Gauss-Markov sources.

Lemma 4 (Second-order lossy AEP for the nonstationary Gauss-Markov sources)**.** *For the Gauss-Markov source with $a > 1$, let $P_{V_1^{\star n}}$ be the random variable that attains the minimum in (48) with X_1^n there replaced by U_1^n. It holds that*

$$\mathbb{P}\left[\log \frac{1}{P_{V_1^{\star n}}\left(\mathcal{B}\left(U_1^n, d\right)\right)} \geq \jmath_{U_1^n}(U_1^n, d) + p(n)\right] \leq \frac{1}{q(n)}, \tag{81}$$

where

$$p(n) \triangleq c_1 (\log n)^{c_2} + c_3 \log n + c_4, \tag{82}$$

$$q(n) \triangleq \Theta(\log n), \tag{83}$$

with positive constants c_i's, $i = 1, ..., 4$.

The proof of Lemma 4 is presented in Appendix F-E below. The proof of Theorem 7 (using the random coding bound (79) and Lemma 4) is presented in Appendix E below.

F. The Connection between Lossy AEP and Parameter Estimation

The proof of lossy AEP in the form of Lemma 4 is technical even for stationary memoryless sources. A lossy AEP for stationary α-mixing processes was derived in [37, Cor. 17]. For stationary memoryless sources with single-letter distribution P_X, the idea in [22, Lem. 2] is to form a typical set $\mathcal{F}_n$ of source outcomes [22, Lem. 4] using the product of the empirical distributions [22, Eq. (270)]: $P_{\hat{X}} \times \ldots \times P_{\hat{X}}$, where $P_{\hat{X}}(x) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = x\}$ is the empirical distribution of a given source sequence x_1^n, and then to show that the inequality inside the bracket in (81) holds for $x_1^n \in \mathcal{F}_n^c$ and that the probability of the complement set $\mathcal{F}_n^c$ is at most $1/q(n)$, where $p(n) = C \log n + c$ and $q(n) = K/\sqrt{n}$ [22, Lem. 2]. The Gauss-Markov source is not memoryless, and it is nonstationary for $a > 1$. To form a typical set of source outcomes, we define the following proxy random variables using the estimator $\hat{a}_{\text{ML}}(u_1^n)$ in (2).

Definition 2 (Proxy random variables)**.** *For each sequence u_1^n of length n generated by the Gauss-Markov source, define the proxy random variable $\hat{X}_1^n$ as an n-dimensional Gaussian random vector with independent coordinates, each of which follows the distribution $\mathcal{N}(0, \hat{\sigma}_{n,i}^2)$ with*

$$\hat{\sigma}_{n,1}^2 \triangleq \sigma^2 \hat{a}_{ML}(u_1^n)^{2n}, \tag{84}$$

$$\hat{\sigma}_{n,i}^2 \triangleq \frac{\sigma^2}{1 + \hat{a}_{ML}(u_1^n)^2 - 2\hat{a}_{ML}(u_1^n)\cos\frac{i\pi}{n+1}}, \quad 2 \leq i \leq n, \tag{85}$$

where $\hat{a}_{ML}(u_1^n)$ is in (2) above.

Remark 6. The proxy random variable in Definition 2 differs from that in [47, Eq. (119)] for the stationary case in the behavior of the largest variance $\hat{\sigma}_{n,1}^2$. For each realization u_1^n, we construct the Gaussian random vector $\hat{X}_1^n$ according to (84)-(85), which is a proxy to the decorrelation X_1^n in (61) above.

The variances of $\hat{X}_i$ and X_i are very close due to the closeness of $\hat{a}_{\text{ML}}(u_1^n)$ to a (Corollary 1).

Remark 7. Since the proxy random variable $\hat{X}_1^n$ depends on the realization of U_1^n, Definition 2 defines the joint distribution of $(X_1^n, \hat{X}_1^n)$, where X_1^n is the decorrelation of U_1^n in (61) above.

The following convex optimization problem will be instrumental: for two generic random vectors A_1^n and B_1^n with distributions $P_{A_1^n}$ and $P_{B_1^n}$, respectively, define

$$\mathbb{R}(A_1^n, B_1^n, d) \triangleq \inf_{\substack{P_{F_1^n|A_1^n}: \\ \mathbb{E}[\mathsf{d}(A_1^n, F_1^n)] \le d}} \frac{1}{n} D(P_{F_1^n|A_1^n} || P_{B_1^n} | P_{A_1^n}), \tag{86}$$

where $D(P_{F_1^n|A_1^n} || P_{B_1^n} | P_{A_1^n})$ is the conditional relative entropy. See Appendix F-B for detailed discussions on this optimization problem.

For each realization u_1^n (equivalently, each $x_1^n = \mathsf{S}' u_1^n$ with the $n \times n$ matrix S defined in the text above (60)), we define n random variables $m_i(u_1^n)$, $i = 1, \ldots, n$ as follows.

- Let X_1^n the decorrelation of U_1^n in (61) above. Let $Y_1^{\star n}$ be the random variable that attains the infimum in $\mathbb{R}_{X_1^n}(d)$.
- For each u_1^n, choose A_1^n in (86) to be the proxy random variable $\hat{X}_1^n$, and let B_1^n to be $Y_1^{\star n}$. Let $\hat{F}_1^{\star n}$ be the random variable that attains the infimum in $\mathbb{R}(\hat{X}_1^n, Y_1^{\star n}, d)$.

Then, for each $i = 1, \ldots, n$, define

$$m_i(u_1^n) \triangleq \mathbb{E}\left[(\hat{F}_i^\star - x_i)^2 | \hat{X}_i = x_i \right]. \tag{87}$$

Denote

$$\eta_n \triangleq \sqrt{\frac{\log\log n}{n}}. \tag{88}$$

The typical set for the Gauss-Markov source is then defined as follows.

Definition 3 (Typical set)**.** *For any $d \in (0, d_{\max})$, $n \ge 2$ and a constant $p > 0$, define $\mathcal{T}(n,p)$ to be the set of vectors $u_1^n \in \mathbb{R}^n$ that satisfy the following conditions:*

$$|\hat{a}_{ML}(u_1^n) - a| \le \eta_n, \tag{89}$$

$$\left| \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i^2}{\sigma_{n,i}^2} \right)^k - (2k-1)!! \right| \le 2, \quad k = 1, 2, 3, \tag{90}$$

$$\left| \frac{1}{n} \sum_{i=1}^n m_i(u_1^n) - d \right| \le p\eta_n, \tag{91}$$

where $x_1^n = \mathsf{S}' u_1^n$ is the decorrelation (61) and $\sigma_{n,i}^2$'s are defined in (60) above.

The typical set in Definition 3 is in the same form as that in the stationary case [47, Def. 2], but the definitions of proxy random variables and the analyses are different.

Theorem 8. *For any $d \in (0, d_{\max})$, there exists a constant $p > 0$ such that for the probability that the Gauss-Markov source produces a typical sequence satisfies*

$$\mathbb{P}\left[U_1^n \in \mathcal{T}(n,p) \right] \ge 1 - \Theta\left(\frac{1}{\log n} \right). \tag{92}$$

Corollary 1 is essential to the proof of Theorem 8. See the details in Appendix F-C.

Let $\mathcal{E}$ denote the event inside the square bracket in (81). Then, we prove Lemma 4 by intersecting $\mathcal{E}$ with the typical set $\mathcal{T}(n,p)$ and the complement $\mathcal{T}(n,p)^c$, respectively, and then bounding the probability of the two intersections separately. See Appendix F-E for the details.

V. Discussion

A. Stationary and Nonstationary Gauss-Markov Processes

It took several decades [13–17] to completely understand the difference in rate-distortion functions between stationary and nonstationary Gaussian autoregressive sources. We briefly summarize this subtle difference here to make the point that generalizing results from the stationary case to the nonstationary one is natural but nontrivial.

Since $\det(\mathsf{F}) = 1$, the eigenvalues $\mu_{n,i}$'s of $\mathsf{F}'\mathsf{F}$ satisfy

$$\prod_{i=1}^n \mu_{n,i} = 1. \tag{93}$$

Using (93), we can equivalently rewrite (50) as

$$\mathbb{R}_{U_1^n}(d) = \frac{1}{n} \sum_{i=1}^n \max\left(0, \ \frac{1}{2} \log \frac{\sigma_{n,i}^2}{\theta_n} \right), \tag{94}$$

where $\theta_n > 0$ is in (51) and $\sigma_{n,i}^2$'s are in (60). Both (50) and (94) are valid expressions for the n-th order rate-distortion function $\mathbb{R}_{U_1^n}(d)$, regardless of whether the source is stationary or nonstationary. The classical Kolmogorov reverse waterfilling result [13, Eq. (18)], obtained by taking the limit in (94), implies that the rate-distortion function of the *stationary* Gauss-Markov source ($0 < a < 1$) is given by (the subscript K stands for Kolmogorov)

$$\mathbb{R}_{\text{K}}(d) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left(0, \ \frac{1}{2} \log \frac{\sigma^2}{\theta g(w)} \right) dw, \tag{95}$$

where $\theta > 0$ is given in (54) and $g(w)$ is given in (55). While (53) and (54) are valid for both stationary and nonstationary cases, Hashimoto and Arimoto [16] noticed in 1980 that (95) is incorrect for the nonstationary Gaussian autoregressive source. The reason is the different asymptotic behaviors of the eigenvalues $\mu_{n,i}$'s of $\mathsf{F}'\mathsf{F}$ (52) in the stationary and nonstationary cases: while in the stationary case, the spectrum is bounded away from zero, in the nonstationary case, the smallest eigenvalue $\mu_{n,1}$ approaches 0, causing a discontinuity. By treating that smallest eigenvalue in a special way, Hashimoto and Arimoto [16, Th. 2] showed that

$$\mathbb{R}_{\text{HA}}(d) = \mathbb{R}_{\text{K}}(d) + \log(\max(a, 1)) \tag{96}$$

is the correct rate-distortion function for both stationary and nonstationary Gauss-Markov sources, where the subscript HA stands for the authors of [16]. For the general higher-order Gaussian autoregressive source, the correction term needed in (96) depends on the unstable roots of the characteristic polynomial of the source, see [16, Th. 2] for the details. In 2008, Gray and Hashimoto [17] showed the equivalence between $\mathbb{R}_{\text{HA}}(d)$ in (96), obtained by taking a limit in (94), and Gray's result $\mathbb{R}_U(d)$ in (53), obtained by taking a limit in (50).

The tool that allows one to take limits in (94) and (50) is the following theorem on the asymptotic eigenvalue distribution of the almost Toeplitz matrix $\mathsf{F}'\mathsf{F}$, which is the (rescaled) inverse of the covariance matrix of U_1^n. Denote

$$\alpha \triangleq \min_{w\in[-\pi,\pi]} \; g(w) = (a-1)^2, \tag{97}$$

and

$$\beta \triangleq \max_{w\in[-\pi,\pi]} \; g(w) = (a+1)^2. \tag{98}$$

Gray [65, Th. 2.4] generalized the result of Grenander and Szegö [66, Th. in Sec. 5.2] on the asymptotic eigenvalue distribution of Toeplitz forms to that of matrices that are asymptotically equivalent to Toeplitz forms, see [65, Chap. 2.3] for the details. Define

$$\alpha' \triangleq \inf_{n\in\mathbb{N},\; i\in[n]} \; \mu_{n,i}. \tag{99}$$

Theorem 9 (Gray [15, Eq. (19)], Hashimoto and Arimoto [16, Th. 1])**.** *For any continuous function $F(t)$ over the interval*

$$t \in [\alpha',\ \beta]\,, \tag{100}$$

the eigenvalues $\mu_{n,i}$'s of $\mathsf{F}'\mathsf{F}$ with F in (52) satisfy

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F(\mu_{n,i}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F\left(g(w)\right)\; dw, \tag{101}$$

where $g(w)$ is defined in (55).

The eigenvalues $\mu_{n,i}$'s behave quite differently in the following three cases, leading the subtle difference in rate-distortion functions.

1) For the stationary case $a \in (0,1)$, it can be easily shown [47, Eq. (71)] that $\alpha' = \alpha > 0$ and all eigenvalues $\mu_{n,i}$'s lie in between α and β. Kolmogorov's formula (95) is obtained by applying Theorem 9 to (94) using the function

$$F_{\mathrm{K}}(t) \triangleq \max\left(0,\ \frac{1}{2}\log\frac{\sigma^2}{\theta t}\right), \tag{102}$$

where $\theta > 0$ is given by (54).

2) For unit-root processes / Wiener processes $a = 1$, closed-form expressions of $\mu_{n,i}$'s are given by Berger [14, Eq. (2)]. Those results imply that the smallest eigenvalue $\mu_{n,1}$ is of order $\Theta\left(\frac{1}{n^2}\right)$ and $\alpha' = \alpha = 0$. Using the same function as in (102), Berger obtained the rate-distortion functions for the Wiener processes $a = 1$ [14, Eq. 4] [1].

3) For the nonstationary case $a > 1$, we have $\alpha' = 0 < \alpha$, the smallest eigenvalue $\mu_{n,1}$ is of order $\Theta(a^{-2n})$ and the other $n-1$ eigenvalues lie in between α and β. This behavior of eigenvalues was shown by Hashimoto and Arimoto [16, Lemma] for higher-order Gaussian autoregressive sources, and we will show a refined version for the Gauss-Markov source in Lemma 5 below. As pointed out by [16, Th. 1], an application of Theorem 9 using the function (102) fails to yield the correct rate-distortion function for nonstationary sources due to the discontinuity of $F_{\mathrm{K}}(t)$ at 0.

Gray [15, Eq. (22)] and Hashimoto and Arimoto [16] circumvent the above difficulty in two different ways, which lead to (53) and (96), respectively. Gray [15] applied Theorem 9 on (50) using the function

$$F_{\mathrm{G}}(t) = \frac{1}{2}\log\max\left(t,\ \frac{\sigma^2}{\theta}\right), \tag{103}$$

which is indeed continuous at 0, while Hashimoto and Arimoto [16, Th. 2] still use the function $F_{\mathrm{K}}(t)$ but consider $\mu_{n,1}$ and $\mu_{n,i},\ i \geq 2$ separately:

$$\frac{1}{n}\sum_{i=2}^{n} F_{\mathrm{K}}(\mu_{n,i}) + \frac{1}{n}F_{\mathrm{K}}(\mu_{n,1}), \tag{104}$$

which in the limit yields (96) by plugging $\mu_{n,1} = \Theta(a^{-2n})$ into (102).

B. New Results on the Spectrum of the Covariance Matrix

The following result on the scaling of the eigenvalues $\mu_{n,i}$'s refines [16, Lemma], and its proof is presented in Appendix B-D.

Lemma 5. *Fix $a > 1$. For any $i = 2,\ldots,n$, the eigenvalues of $\mathsf{F}'\mathsf{F}$ are bounded as*

$$\xi_{n-1,i-1} \leq \mu_{n,i} \leq \xi_{n,i}, \tag{105}$$

where

$$\xi_{n,i} \triangleq 1 + a^2 - 2a\cos\left(\frac{i\pi}{n+1}\right). \tag{106}$$

The smallest eigenvalue is bounded as

$$2\log a + \frac{c_2}{n} \geq -\frac{1}{n}\log\mu_{n,1} \geq 2\log a - \frac{c_1}{n}, \tag{107}$$

where $c_1 > 0$ and c_2 are constants given by

$$c_1 = 2\log(a+1) + \frac{a\pi}{a^2-1}, \tag{108}$$

$$c_2 = 2\log\frac{a}{a^2-1} + \frac{2a\pi}{a^2-1}. \tag{109}$$

Remark 8. The constant c_1 in (108) is positive, while c_2 in (109) can be positive, zero or negative, depending on the value of $a > 1$. Lemma 5 indicates that a^{-2n} is a good approximation to $\mu_{n,1}$. Using (105)–(106), we deduce that for $i = 2,\ldots,n$,

$$\mu_{n,i} \in [\alpha,\beta]. \tag{110}$$

Based on Lemma 5, we obtain a nonasymptotic version of Theorem 9, which is useful in the analysis of the dispersion, in particular, in deriving Proposition 1 in Appendix C-A below.

Theorem 10. *Fix any $a > 1$. For any bounded, L-Lipschitz and nondecreasing function (or nonincreasing function) $F(t)$ over the interval (100) and any $n \geq 1$, the eigenvalues $\mu_{n,i}$'s of $\mathsf{F}'\mathsf{F}$ satisfy*

$$\left|\frac{1}{n}\sum_{i=1}^{n} F(\mu_{n,i}) - \frac{1}{2\pi}\int_{-\pi}^{\pi} F\left(g(w)\right)\; dw\right| \leq \frac{C_L}{n}, \tag{111}$$

[1] To be precise, although the rate-distortion function for the Wiener process is correct in [14, Eq. 4], the proof there is not rigorous since in this case $\alpha' = \alpha = 0$ but $F_{\mathrm{K}}(t)$ is not continuous at $t = 0$ as pointed out in [17, Eq. (23)]. Therefore, the limit leading to [14, Eq. 4] needs extra justifications.

where $g(w)$ is defined in (55) and $C_L > 0$ is a constant that depends on L and the maximum absolute value of F.

The proof of Theorem 10 is in Appendix B-E.

VI. Conclusion

In this paper, we obtain nonasymptotic (Theorem 1) and asymptotic (Theorem 2) bounds on the estimation error of the maximum likelihood estimator of the parameter a of the nonstationary scalar Gauss-Markov process. Numerical simulations in Fig. 1 confirm the tightness of our estimation error bounds compared to previous works. As an application of the estimation error bound (Corollary 1), we find the dispersion for lossy compression of the nonstationary Gauss-Markov sources (Theorems 6 and 7). Future research directions include generalizing the error exponent bounds in this paper, applicable to identification of scalar dynamical systems, to vector systems, and finding the dispersion of the Wiener process.

Appendix A

A. Proof of Theorem 1

Proof. We present the proof of (24). The proof of (25) is similar and is omitted. For any $n \geq 2$, denote by $\mathcal{F}_n$ the σ-algebra generated by $Z_1, \ldots, Z_n$. For any $s > 0$, $\eta > 0$, and $n \geq 2$, we denote W_n the following random variable

$$W_n \triangleq \exp\left\{ s \sum_{i=1}^{n-1} (U_i Z_{i+1} - \eta U_i^2) \right\}. \tag{112}$$

By the Chernoff bound, we have

$$\mathbb{P}\left[\hat{a}_{\mathrm{ML}}(U_1^n) - a \geq \eta\right] \leq \inf_{s>0} \mathbb{E}\left[W_n\right]. \tag{113}$$

To compute $\mathbb{E}[W_n]$, we first condition on $\mathcal{F}_{n-1}$. Since Z_n is the only term in W_n that does not belong to $\mathcal{F}_{n-1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[W_n\right] \\
=&\mathbb{E}\left\{W_{n-1} \cdot \mathbb{E}[\exp(s(U_{n-1}Z_n - \eta U_{n-1}^2))|\mathcal{F}_{n-1}]\right\} &(114)\\
=&\mathbb{E}\left[W_{n-1} \cdot \exp(\alpha_1 U_{n-1}^2)\right], &(115)
\end{aligned}$$

where α_1 is the deterministic function of s and η defined in (18), and (115) follows from the moment generating function of Z_n. To obtain a recursion, we condition on $\mathcal{F}_{n-2}$. Since U_{n-1}^2 and $U_{n-2}Z_{n-1}$ are the only two terms in $W_{n-1} \cdot \exp(\alpha_1 U_{n-1}^2)$ that do not belong to $\mathcal{F}_{n-2}$, we use the relation $U_{n-1} = aU_{n-2} + Z_{n-1}$ and we complete squares in Z_{n-1} to obtain

$$\begin{aligned}
&W_{n-1} \cdot \exp(\alpha_1 U_{n-1}^2) \\
=&W_{n-2} \cdot \exp\left\{\alpha_1 \left(Z_{n-1} + (a + \frac{s}{2\alpha_1})U_{n-2}\right)^2 + \right. \\
&\left. (a^2\alpha_1 - s\eta)U_{n-2}^2 - \alpha_1 \left(a + \frac{s}{2\alpha_1}\right)^2 U_{n-2}^2 \right\}. &(116)
\end{aligned}$$

Furthermore, using the formula for the moment generating function of the noncentral χ^2-distributed random variable

$$\left(Z_{n-1} + \left(a + \frac{s}{2\alpha_1}\right) U_{n-2}\right)^2 \tag{117}$$

with 1 degree of freedom, we obtain

$$\begin{aligned}
&\mathbb{E}\left[W_{n-1} \cdot \exp(\alpha_1 U_{n-1}^2)\right] \\
=&\frac{1}{\sqrt{1 - 2\sigma^2\alpha_1}} \mathbb{E}\left[W_{n-2} \cdot \exp(\alpha_2 U_{n-2}^2)\right]. &(118)
\end{aligned}$$

This is where our method diverges from Rantzer [10, Lem. 5], who chooses $s = \frac{\eta}{\sigma^2}$ and bounds $\alpha_2 \leq \alpha_1$ (due to Property A4 in Appendix A-B below) in (118). Instead, by conditioning on $\mathcal{F}_{n-3}$ in (118) and repeating the above recursion for another $n-2$ times, we compute $\mathbb{E}[W_n]$ exactly using the sequence $\{\alpha_\ell\}$:

$$\mathbb{E}\left[W_n\right] = \exp\left\{ -\frac{1}{2} \sum_{\ell=1}^{n-1} \log(1 - 2\sigma^2\alpha_\ell) \right\}. \tag{119}$$

If $s \notin \mathcal{S}_n^+$, then by the definition of the set $\mathcal{S}_n^+$ we have $\mathbb{E}[W_n] = +\infty$. Therefore,

$$\inf_{s>0} \mathbb{E}\left[W_n\right] = \inf_{s \in \mathcal{S}_n^+} \mathbb{E}\left[W_n\right]. \tag{120}$$

□

B. Properties of the Sequence α_ℓ

We derive several important elementary properties about the sequences α_ℓ and β_ℓ. First, we consider α_ℓ. We find the two fixed points $r_1 < r_2$ of the recursive relation (19) by solving the following quadratic equation in x:

$$2\sigma^2 x^2 + [a^2 + 2\sigma^2 s(a + \eta) - 1]x + \alpha_1 = 0. \tag{121}$$

Property A1: For any $s > 0$ and $\eta > 0$, (121) has two roots $r_1 < r_2$, and $r_1 < 0$. The two roots r_1 and r_2 are given by

$$r_1 = \frac{-[a^2 + 2\sigma^2(a+\eta)s - 1] - \sqrt{\Delta}}{4\sigma^2}, \tag{122}$$

$$r_2 = \frac{-[a^2 + 2\sigma^2(a+\eta)s - 1] + \sqrt{\Delta}}{4\sigma^2}, \tag{123}$$

where Δ denotes the discriminant of (121):

$$\begin{aligned}
\Delta = &4\sigma^4[(a+\eta)^2 - 1]s^2 + \\
&4\sigma^2[(a+\eta)(a^2-1) + 2\eta]s + (a^2-1)^2. &(124)
\end{aligned}$$

Proof. Note that the discriminant Δ satisfies

$$\Delta > (a^2 - 1)^2 > 0, \tag{125}$$

where we used $a > 1$. Then, (122) implies $r_1 < 0$. □

Property A2: For $\frac{2\eta}{\sigma^2} \neq s > 0$ and $\eta > 0$, the sequence $\frac{\alpha_\ell - r_1}{\alpha_\ell - r_2}$ is a geometric sequence with common ratio

$$q \triangleq \frac{[a^2 + 2\sigma^2 s(a+\eta)] + 2\sigma^2 r_1}{[a^2 + 2\sigma^2 s(a+\eta)] + 2\sigma^2 r_2}. \tag{126}$$

Furthermore,

$$q \in (0, 1), \tag{127}$$

and it follows immediately that

$$\alpha_\ell = r_1 + \frac{(r_1 - r_2)\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell-1}}{1 - \frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell-1}}, \tag{128}$$

$$= r_2 + \frac{r_2 - r_1}{\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell-1} - 1}. \tag{129}$$

Proof. Using the recursion (19) and the fact that r_1 and r_2 are the fixed points of (19), one can verify that $\frac{\alpha_\ell - r_1}{\alpha_\ell - r_2}$ is a geometric sequence with common ratio q given by (126). The relation (127) is verified by direct computations using (122) and (123). □

Property A3: For any $\frac{2\eta}{\sigma^2} \neq s > 0$ and $\eta > 0$, we have

$$\lim_{\ell \to \infty} \alpha_\ell = r_1. \tag{130}$$

For $s = \frac{2\eta}{\sigma^2}$, we have $\alpha_\ell = 0 = r_2 > r_1, \ \forall \ell \geq 1$.

Proof. The limit (130) follows from (127) and (128). Plugging $s = \frac{2\eta}{\sigma^2}$ into (18) yields $\alpha_1 = 0$, which implies by (19) that $\alpha_\ell = 0$ for $\ell \geq 1$. □

Property A4: For any $s \in \mathcal{I}_\eta$, we have $\alpha_\ell < 0$ and α_ℓ decreases to r_1 geometrically. For $s > \frac{2\eta}{\sigma^2}$, (130) still holds, but the convergence is not monotone: there exists an $\ell^\star \geq 1$ such that $\alpha_\ell > 0$ and increases to $\alpha_{\ell^\star}$ for $1 \leq \ell \leq \ell^\star$; and $\alpha_\ell < 0$ and increases to r_1 for $\ell > \ell^\star$.

Proof. Due to (129), the monotonicity of α_ℓ depends on the signs of $r_2 - r_1$ and $\frac{\alpha_1 - r_1}{\alpha_1 - r_2}$. Note that $r_2 - r_1 > 0$ by Property A1. Plugging $x = \alpha_1$ into (121), we have

$$(\alpha_1 - r_1)(\alpha_1 - r_2) = (a + \sigma^2 s)^2 \alpha_1. \tag{131}$$

Since for $s \in \mathcal{I}_\eta$ we have $\alpha_1 < 0$ by (18), (131) implies that $\frac{\alpha_1 - r_1}{\alpha_1 - r_2} < 0$ for any $s \in \mathcal{I}_\eta$. This immediately implies that α_ℓ decreases to r_1 due to (128) and (129). Therefore, $\alpha_\ell \leq \alpha_1 < 0, \ \forall \ell \geq 1$. For any $s > \frac{2\eta}{\sigma^2}$, we have $\alpha_1 > 0$ and $\frac{\alpha_1 - r_1}{\alpha_1 - r_2} > 0$. In fact, since $r_1 < 0$, we have $\alpha_1 > r_2$, which implies $\frac{\alpha_1 - r_1}{\alpha_1 - r_2} > 1$. Therefore, the conclusion follows from (129). □

Property A5: For any $\eta > 0$, the root r_1 in (122) is a decreasing function in $s > 0$.

Proof. Direct computations using (122), (124) and the assumption that $a > 1$. □

C. Properties of the Sequence β_ℓ

The sequence β_ℓ is analyzed similarly, although it is slightly more involved than α_ℓ. We only consider $0 < s \leq \frac{2\eta}{\sigma^2}$ in the rest of this section. We find the two fixed points $t_1 < t_2$ of the recursive relation (21) by solving the following quadratic equation in x:

$$2\sigma^2 x^2 + [a^2 + 2\sigma^2 s(-a + \eta) - 1]x + \beta_1 = 0. \tag{132}$$

Property B1: For $s = \frac{2\eta}{\sigma^2}$, we have $\beta_\ell = 0, \ \forall \ell \geq 1$. For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, (132) has two distinct roots $t_1 < 0 < t_2$, given by

$$t_1 = \frac{-[a^2 + 2\sigma^2 s(-a + \eta) - 1] - \sqrt{\Gamma}}{4\sigma^2}, \tag{133}$$

$$t_2 = \frac{-[a^2 + 2\sigma^2 s(-a + \eta) - 1] + \sqrt{\Gamma}}{4\sigma^2}, \tag{134}$$

where the discriminant Γ of (132) is

$$\Gamma = 4\sigma^4[(-a + \eta)^2 - 1]s^2 + 4\sigma^2[(-a + \eta)(a^2 - 1) + 2\eta]s + (a^2 - 1)^2. \tag{135}$$

Proof. We verify that $\Gamma > 0$ for any $\eta > 0$ and $s \in \mathcal{I}_\eta$. The reason that $\Gamma > 0$ is not as obvious as (125) is due to the subtle difference between (124) and (135) in the negative sign of a. Note that Γ in (135) is a quadratic equation in s and the discriminant of Γ is given by (with some elementary manipulations)

$$\gamma = 16\sigma^4(2a\eta - a^2 + 1)^2 \geq 0. \tag{136}$$

Hence, in general, (135) has two roots (distinct when $\eta \neq \frac{a^2 - 1}{2a}$) and Γ could be positive or negative. However, an analysis of two cases $(-a + \eta)^2 - 1 \geq 0$ and $(-a + \eta)^2 - 1 < 0$ reveals that $\Gamma > 0$ for any $\eta > 0$ and $s \in \mathcal{I}_\eta$. Therefore, (132) has two distinct roots $t_1 < t_2$ given in (133) and (134) above. From (132), we have $t_1 t_2 = \frac{\beta_1}{2\sigma^2}$, which is negative for $s \in \mathcal{I}_\eta$. Therefore, we have $t_1 < 0 < t_2$. □

Property B2: For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, the sequence $\frac{\beta_\ell - t_1}{\beta_\ell - t_2}$ is a geometric sequence with common ratio

$$p \triangleq \frac{[a^2 + 2\sigma^2 s(-a + \eta)] + 2\sigma^2 t_1}{[a^2 + 2\sigma^2 s(-a + \eta)] + 2\sigma^2 t_2}. \tag{137}$$

In addition, for any $\eta > 0$ and $s \in \mathcal{I}_\eta$, we also have

$$p \in (0, 1). \tag{138}$$

It follows immediately that

$$\beta_\ell = t_1 + \frac{(t_1 - t_2)\frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell - 1}}{1 - \frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell - 1}}, \tag{139}$$

$$= t_2 + \frac{t_2 - t_1}{\frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell - 1} - 1}. \tag{140}$$

Proof. Similar to that of Property A2 above for α_ℓ. □

Property B3: For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, we have $\beta_\ell \leq \beta_1 < 0$ and β_ℓ decreases to t_1 geometrically:

$$\lim_{\ell \to \infty} \beta_\ell = t_1. \tag{141}$$

Proof. This can be verified using (139) and (140) by noticing that $t_2 - t_1 > 0$ and for $s \in \mathcal{I}_\eta$,

$$(\beta_1 - t_1)(\beta_2 - t_2) = (a - \sigma^2 s)^2 \beta_1 < 0. \tag{142}$$

□

Property B4: For any constant $a > 1$, recall the two thresholds η_1 and η_2, defined in (37) and (38) in Section III-A above, respectively. Then,

1) When $0 < \eta \leq \eta_1$, the root t_1 in (133) is an increasing function in $s \in \mathcal{I}_\eta$.
2) When $\eta \geq \eta_2$, t_1 is a decreasing function in $s \in \mathcal{I}_\eta$.
3) When $\eta_1 < \eta < \eta_2$, t_1 is a decreasing function in $s \in (0, s^\star)$; and an increasing function in $s \in \left(s^\star, \frac{2\eta}{\sigma^2}\right)$, where $s^\star$ is the unique solution in the interval $\mathcal{I}_\eta$ to

$$\left.\frac{dt_1}{ds}\right|_{s = s^\star} = 0, \tag{143}$$

and $s^\star$ is given by

$$s^\star \triangleq \frac{a\eta(\eta - \eta_1)}{\sigma^2(1 - (\eta - a)^2)}. \tag{144}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

Proof. Using (133) and (135), we compute the derivatives of t_1 as follows:

$$\frac{dt_1}{ds} = -\frac{\eta - a}{2} - \frac{1}{\sqrt{\Gamma}}\Bigg\{\sigma^2[(-a+\eta)^2 - 1]s + \frac{1}{2}[(-a+\eta)(a^2-1) + 2\eta]\Bigg\}, \tag{145}$$

$$\frac{d^2t_1}{ds^2} = \frac{\sigma^2(2a\eta - a^2 + 1)^2}{\Gamma^{\frac{3}{2}}} \geq 0. \tag{146}$$

To simplify notations, denote by $L(s)$ the first derivative:

$$L(s) \triangleq \frac{dt_1}{ds}(s). \tag{147}$$

From (145), we have

$$L(0) = \frac{-a^2(\eta - \eta_1)}{a^2 - 1}, \tag{148}$$

and

$$L\left(\frac{2\eta}{\sigma^2}\right) = \begin{cases} \frac{-2(2\eta - a)(\eta - \eta_2)(\eta - \eta_2')}{(a-2\eta)^2 - 1}, & \eta \in \left(0, \frac{a-1}{2}\right) \cup \left(\frac{a+1}{2}, +\infty\right) \\ \frac{\eta}{1-(a-2\eta)^2}, & \eta \in \left(\frac{a-1}{2}, \frac{a+1}{2}\right), \end{cases} \tag{149}$$

where η_2' is given by

$$\eta_2' \triangleq \frac{3a - \sqrt{a^2+8}}{4}. \tag{150}$$

Since $L(s)$ is an increasing function in s due to (146), to determine the monotonicity of t_1, we only need to consider the following three cases.

a) When $L(0) \geq 0$, or equivalently, $0 < \eta \leq \eta_1$, we have $L(s) \geq 0$ for any $s \in \mathcal{I}_\eta$. Hence, t_1 is an increasing function in s.

b) When $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$, we have $L(s) \leq 0$ for any $s \in \mathcal{I}_\eta$. Hence, t_1 is a decreasing function in s. We now show that $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is equivalent to $\eta \geq \eta_2$. When $\eta \in \left(\frac{a-1}{2}, \frac{a+1}{2}\right)$, we have $L\left(\frac{2\eta}{\sigma^2}\right) > 0$ by (149) and $\eta > 0$. When $\eta \in \left(0, \frac{a-1}{2}\right) \cup \left(\frac{a+1}{2}, +\infty\right)$, it is easy to see from (149) that $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is equivalent to $\eta \in [\eta_2', a/2] \cup [\eta_2, +\infty)$. Hence, the equivalent condition for $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is $\eta \in [\eta_2, +\infty)$.

c) When $L(0) < 0$ and $L\left(\frac{2\eta}{\sigma^2}\right) > 0$, or equivalently, $\eta \in (\eta_1, \eta_2)$, solving (143) using (145) yields (144). Since $L(s)$ is monotonically increasing due to (146), we know that $s^\star$ given by (144) is the unique solution to (143) in $\mathcal{I}_\eta$, and $L(s) \leq 0$ for $s \in (0, s^\star]$ and $L(s) > 0$ for $s \in (s^\star, 2\eta/\sigma^2)$. □

D. Proof of Lemma 1

Proof. We first show the monotone decreasing property. The set $\mathcal{S}_{n+1}^+$ contains all $s > 0$ such that $a_1, ..., a_n, a_{n+1}$ are all less than $1/2\sigma^2$, while the set $\mathcal{S}_n^+$ contains all $s > 0$ such that $a_1, ..., a_n$ are all less than $1/2\sigma^2$, hence $\mathcal{S}_{n+1}^+ \subseteq \mathcal{S}_n^+$. The same argument yields the conclusion for $\mathcal{S}_n^-$.

We then prove that $\mathcal{S}_\infty^+ = \left(0, 2\eta/\sigma^2\right]$. Property A4 above in Appendix A-B implies that for any $0 < s \leq 2\eta/\sigma^2$, we have $\alpha_\ell \leq 0 < \frac{1}{2\sigma^2}$. Hence $\left(0, 2\eta/\sigma^2\right] \subseteq \mathcal{S}_n^+$ for any $n \geq 1$. To show the other direction, it suffices to show that for any $s > \frac{2\eta}{\sigma^2}$, there exists $n \in \mathbb{N}$ such that $\alpha_n \geq \frac{1}{2\sigma^2}$. Let $\ell^\star$ be the integer defined in Property A4 above. Then, $\ell^\star$ satisfies the following two conditions

$$\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell^\star - 1} \geq 1, \tag{151}$$

$$\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell^\star} < 1. \tag{152}$$

We show that $\alpha_{\ell^\star} \geq \frac{1}{2\sigma^2}$, which would complete the proof. Due to $r_2 - r_1 > 0$, using (129) and (152), we have

$$\alpha_{\ell^\star} \geq r_2 + \frac{r_2 - r_1}{\frac{1}{q} - 1} \tag{153}$$

$$= \frac{r_2 - r_1 q}{1 - q} \tag{154}$$

$$= \frac{1}{2\sigma^2}, \tag{155}$$

where (155) [2] is by plugging (122), (123) and (126) into (154).

Finally, to show (31), for any $0 < s \leq 2\eta/\sigma^2$, we have $\beta_\ell \leq 0 < \frac{1}{2\sigma^2}$, $\forall \ell \geq 1$, hence $\left(0, 2\eta/\sigma^2\right] \subseteq \mathcal{S}_\infty^-$. The other direction cannot hold since there are many counterexamples, e.g., $a = 1.2$, $\sigma^2 = 1$, $\eta = 0.15$ and $s = 0.35 > \frac{2\eta}{\sigma^2}$, where the sequence β_ℓ increases monotonically to $t_1 \approx 0.0411 < \frac{1}{2\sigma^2}$. Hence, in this case, $0.35 \in \mathcal{S}_\infty^-$ but $0.35 \notin \left(0, \frac{2\eta}{\sigma^2}\right]$. □

E. Proof of Theorem 2

Proof. Theorem 1 and Lemma 1 imply that for any $s \in \mathcal{I}_\eta$,

$$\liminf_{n\to\infty} P^+(n, a, \eta) \geq \lim_{n\to\infty} \frac{1}{2n} \sum_{\ell=1}^{n-1} \log(1 - 2\sigma^2 \alpha_\ell). \tag{156}$$

Recall that α_ℓ depends on s. By (130), the continuity of the function $x \mapsto \log(1-x)$ and the Cesàro mean convergence, we have

$$\liminf_{n\to\infty} P^+(n, a, \eta) \geq \frac{1}{2} \log(1 - 2\sigma^2 r_1), \tag{157}$$

where r_1 depends on s via (122). Since (157) holds for any $s \in \mathcal{I}_\eta$, using Property A5 in Appendix A-B above and supremizing (157) over $s \in \mathcal{I}_\eta$, we obtain (33). Specifically, the supremum of (157) over $s \in \mathcal{I}_\eta$ is achieved in the limit of s going to the right end point $2\eta/\sigma^2$. Plugging $s = 2\eta/\sigma^2$ into (122), we obtain the corresponding value for r_1:

$$-\frac{(a+2\eta)^2 - 1}{2\sigma^2}, \tag{158}$$

which is further substituted into (157) to yield (33).

Similarly, to show (34), using Property B3 in Appendix A-C above, we have

$$\liminf_{n\to\infty} P^-(n, a, \eta) \geq \sup_{s \in \mathcal{I}_\eta} \frac{1}{2} \log(1 - 2\sigma^2 t_1). \tag{159}$$

Then, by Property B4 in Appendix A-C above, the supermizer s' in (159) is given by

$$s' = \begin{cases} 0, & 0 < \eta \leq \eta_1 \\ s^\star, & \eta_1 < \eta < \eta_2 \\ \frac{2\eta}{\sigma^2}, & \eta \geq \eta_2, \end{cases} \tag{160}$$

[2] It is pretty amazing that (155) is in fact an equality.

where $s^\star$ is given by (144). Plugging (160) into (159) yields (34).

Finally, the bound (35) follows from (33) and (34), since

$$\begin{aligned}&\mathbb{P}\left[|\hat{a}_{\mathrm{ML}}(U_1^n)-a|>\eta\right]\\ =&\mathbb{P}\left[(\hat{a}_{\mathrm{ML}}(U_1^n)-a)>\eta\right]+\mathbb{P}\left[(\hat{a}_{\mathrm{ML}}(U_1^n)-a)<-\eta\right] \quad (161)\end{aligned}$$

and

$$\begin{aligned}&\liminf_{n\to\infty} P(n,a,\eta)\\ =&\liminf_{n\to\infty}\min\left\{P^+(n,a,\eta),P^-(n,a,\eta)\right\} \quad (162)\\ \geq& I^-(a,\eta). \quad (163)\end{aligned}$$

□

F. Proof of Theorem 3

Proof. For any sequence η_n, the proof of Theorem 1 in Appendix A-A above remains valid with α_ℓ replaced by $\alpha_{n,\ell}$ defined in (40) in Section III-C above. We present the proof of (42), and omit that of (43), which is similar. In this regime, for each $n\geq 1$, the proof of Lemma 1 implies that

$$\left(0,\frac{2\eta_n}{\sigma^2}\right)\subseteq\mathcal{S}_n^+. \quad (164)$$

Then, in (24), we choose

$$s=s_n=\frac{\eta_n}{\sigma^2}\in\mathcal{S}_n^+. \quad (165)$$

First, using (122)-(123), (126) and the choice (165), we can determine the asymptotic behavior of quantities involved in determining $\alpha_{n,\ell}$ in (128) and (129) (with η replaced by η_n and s replaced by s_n), summarized in TABLE I.

α_1	r_1	r_2	r_2-r_1	q	$-\frac{\alpha_1-r_1}{\alpha_1-r_2}$
$-\Theta(\eta_n^2)$	$-\Theta(1)$	$\Theta(\eta_n^2)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1/\eta_n^2)$

TABLE I: Order dependence in η_n of the quantities involved in determining $\alpha_{n,\ell}$ in (128) and (129).

We make two remarks before proceeding further. It can be easily verified from (126) that the common ratio q is a constant belonging to $(0,1)$ and

$$\lim_{\eta_n\to 0} q=\frac{1}{a^2}\in(0,1). \quad (166)$$

Hence, for all large n, q is bounded by positive constants between 0 and 1. Besides, from (122), we have

$$\lim_{\eta_n\to 0} r_1=-\frac{a^2-1}{2\sigma^2}. \quad (167)$$

Second, from (128), (24) and the choice (165), we have

$$\begin{aligned}&P^+(n,a,\eta_n)\\ \geq&\frac{n-1}{2n}\log(1-2\sigma^2 r_1)+\\ &\frac{1}{2n}\sum_{\ell=1}^{n-1}\log\left(1-\frac{2\sigma^2(r_2-r_1)}{1-2\sigma^2 r_1}\cdot\frac{\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right)q^{\ell-1}}{1+\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right)q^{\ell-1}}\right), \quad (168)\end{aligned}$$

where r_1,r_2 and q in this regime depend on η_n with order dependence given in TABLE I above. Using the inequality $\log(1-x)\geq\frac{x}{x-1}$, $\forall x\in(0,1)$, we have

$$\begin{aligned}&P^+(n,a,\eta_n)\\ \geq&\frac{n-1}{2n}\log(1-2\sigma^2 r_1)+\\ &\frac{1}{2n}\sum_{\ell=1}^{n-1}\frac{-1}{\frac{1-2\sigma^2 r_2}{2\sigma^2(r_2-r_1)}+\frac{1-2\sigma^2 r_1}{2\sigma^2(r_2-r_1)}\cdot\frac{1}{\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right)q^{\ell-1}}}. \quad (169)\end{aligned}$$

Since $1-2\sigma^2 r_2>0$ due to (123), we can further bound $P^+(n,a,\eta_n)$ as

$$\begin{aligned}&P^+(n,a,\eta_n)\\ \geq&\frac{n-1}{2n}\log(1-2\sigma^2 r_1)-\\ &\frac{1}{n}\left(\sum_{\ell=1}^{n-1}q^{\ell-1}\right)\frac{2\sigma^2(r_2-r_1)}{1-2\sigma^2 r_1}\cdot\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right) \quad (170)\\ \geq&\frac{n-1}{2n}\log(1-2\sigma^2 r_1)-\\ &\frac{1}{n}\frac{2\sigma^2(r_2-r_1)}{(1-2\sigma^2 r_1)(1-q)}\cdot\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right) \quad (171)\\ =&\frac{n-1}{2n}\log(1-2\sigma^2 r_1)-\frac{1}{n\Theta(\eta_n^2)}, \quad (172)\end{aligned}$$

where in the last step we used the results in TABLE I. Due to the assumption (41) on η_n and (167), we obtain (42). □

G. Proof of Theorem 4

Proof. We point out the proof changes in generalizing our results to the sub-Gaussian case. There are two changes to be made in the proof of Theorem 1 in Appendix A-A above, the equality from (114) to (115) is replaced by $\leq$ since Z_n is σ-sub-Gaussian; the equality in (118) is replaced by $\leq$ due to Lemma 2. The rest of the proof for Theorem 1 remains the same for the sub-Gaussian case. Since Lemma 1 and Theorem 2, 3 depend only on the properties of the sequences α_ℓ and β_ℓ and not on the distribution of Z_n's as long as Theorem 1 holds, their proofs remain exactly the same for the sub-Gaussian case. □

Appendix B

A. Proof of Lemma 3

Proof. In view of (62), we take the variances on both sides of (59) to obtain

$$\mathbb{V}_U(d)=\limsup_{n\to\infty}\frac{1}{2n}\sum_{i=1}^{n}\min\left[1,\left(\frac{\sigma_{n,i}^2}{\theta_n}\right)^2\right]. \quad (173)$$

Note that $\lim_{n\to\infty}\theta_n=\theta$, where $\theta>0$ is the water level given by (54). Applying Theorem 9 in Section V-A to (173) with the function

$$F(t)\triangleq\frac{1}{2}\min\left[1,\left(\frac{\sigma^2}{\theta t}\right)^2\right], \quad (174)$$

which is continuous at $t=0$, we obtain (65). □

B. An Integral

We present the computation of an interesting integral that is useful to obtain the value of $\mathbb{R}_U(d_{\max})$.

Lemma 6. *For any constant $r \in [-1, 1]$, it holds that*

$$\int_{-\pi}^{\pi} \log(1 - r\cos(w))\ dw = 4\pi \log \frac{\sqrt{1+r} + \sqrt{1-r}}{2}. \tag{175}$$

Proof. Denote

$$I(r) \triangleq \int_{-\pi}^{\pi} \log(1 - r\cos(w))\ dw. \tag{176}$$

By Leibniz's rule for differentiation under the integral sign, we have

$$\frac{\mathrm{d}I(r)}{\mathrm{d}r} = \int_{-\pi}^{\pi} \frac{\partial}{\partial r} \log(1 - r\cos(w))\ dw \tag{177}$$

$$= -2 \cdot \int_0^{\pi} \frac{\cos w}{1 - r\cos w}\ dw. \tag{178}$$

With the change of variable $u = \tan(w/2)$ and partial-fraction decomposition, we obtain the closed-form solution to the integral in (178):

$$\frac{\mathrm{d}I(r)}{\mathrm{d}r} = \frac{2\pi}{r} - \frac{2\pi}{r\sqrt{1-r^2}}. \tag{179}$$

It can be easily verified by directly taking derivatives that the right-side of (175) is indeed the antiderivative of (179). □

C. Derivation of $\mathbb{R}_U(d_{\max})$ in (74)

We present two ways to obtain (74). The first one is to directly use (96) in Section V-A. For $\theta = \theta_{\max}$, we have $\mathbb{R}_{\mathrm{K}}(d_{\max}) = 0$ in (95), then (74) immediately follows from (96). The second method relies on (53). For $\theta = \theta_{\max}$, observe from (53) that

$$\mathbb{R}_U(d_{\max}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log(g(w))\ dw. \tag{180}$$

Then, computing the integral (180) using Lemma 6 in Appendix B-B yields (74).

D. Proof of Lemma 5

Proof. The bound (105) is obtained by partitioning $\mathsf{F}'\mathsf{F}$ into its leading principal submatrix of order $n-1$ and then applying the Cauchy interlacing theorem to that partition, see [47, Lem. 1] for details. To obtain (107), observe from (93)

$$\mu_{n,1} = \left(\prod_{i=2}^{n} \mu_{n,i} \right)^{-1}. \tag{181}$$

Combining (181) and (105) yields

$$L_n \geq -\frac{1}{n} \log \mu_{n,1} \geq R_n, \tag{182}$$

where

$$L_n \triangleq \frac{1}{n} \sum_{i=2}^{n} \log \xi_{n,i} \quad \text{and} \quad R_n \triangleq \frac{1}{n} \sum_{i=1}^{n-1} \log \xi_{n-1,i}. \tag{183}$$

Plugging (106) into (183) and then taking the limit, we obtain

$$\lim_{n\to\infty} L_n = \lim_{n\to\infty} R_n$$

$$= \frac{1}{\pi} \int_0^{\pi} \log(1 + a^2 - 2a\cos(w))\ dw \tag{184}$$

$$= 2\log a, \tag{185}$$

where the last equality is due to Lemma 6 in Appendix B-B above. In the rest of the proof, we obtain the following refinement of (185): for any $n \geq 1$,

$$R_n \geq 2\log a - \frac{c_1}{n}, \tag{186}$$

$$L_n \leq 2\log a + \frac{c_2}{n}, \tag{187}$$

where c_1 and c_2 are the constants given by (108) and (109) in Lemma 5, respectively. Then, (107) will follow directly from (182), (186) and (187).

The proofs of the refinements (186) and (187) are similar, and both are based on the elementary relations between Riemann sums and their corresponding integrals. We present the proof of (186), and omit that of (187). Note that the function $h(w) \triangleq \frac{1}{\pi}\log(1 + a^2 - 2a\cos(w))$ is an increasing function in $w \in [0, \pi]$, and its derivative is bounded above by $M_1 \triangleq \frac{2a}{\pi(a^2-1)}$ for any fixed $a > 1$. Therefore, from (106) and (183), we have

$$\left| R_n + \frac{1}{n}\log(a+1)^2 - \frac{1}{\pi}\int_0^{\pi} \log(g(w))\ dw \right| \leq \frac{M_1 \pi^2}{2n}, \tag{188}$$

and (186) follows immediately. □

E. Proof of Theorem 10

Proof. From Lemma 5, we know that $\alpha' = 0 < \alpha$ (recall (97) and (99)). Since $g(w)$ is an even function, we have

$$I \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} F(g(w))\ dw \tag{189}$$

$$= \frac{1}{\pi} \int_0^{\pi} F(g(w))\ dw. \tag{190}$$

Denote the maximum absolute value of F over the interval (100) by $T > 0$. It is easy to check that the function $F(g(w))$ is $2aL$-Lipschitz since $F(\cdot)$ is L-Lipschitz and the derivative of $g(w)$ is bounded by $2a$. For the following Riemann sum

$$S_n \triangleq \frac{1}{n} \sum_{i=1}^{n} F\left(g\left(\frac{i\pi}{n} \right) \right), \tag{191}$$

the Lipschitz property implies that

$$|S_n - I| \leq \frac{2aL}{\pi n}. \tag{192}$$

For $i \geq 2$, rewrite (106) and (105) as

$$g\left(\frac{(i-1)\pi}{n} \right) \leq \mu_{n,i} \leq g\left(\frac{i\pi}{n+1} \right). \tag{193}$$

Denote the sum in (111) as

$$Q_n \triangleq \frac{1}{n} \sum_{i=1}^{n} F(\mu_{n,i}). \tag{194}$$

Then, separating $F(\mu_{n,1})$ from Q_n and applying (193), we have

$$Q_n \geq S_n - \frac{2T}{n}, \tag{195}$$

$$Q_n \leq \frac{n+1}{n} S_{n+1} + \frac{3T}{n}. \tag{196}$$

Therefore, there is a constant $C_L > 0$ depending on L and T such that (111) holds. □

APPENDIX C

We gather the frequently used notations in this section as follows. For any given distortion threshold $d > 0$,

- let $\theta > 0$ be the water level corresponding to d in the limiting reverse waterfilling (54);
- for each $n \geq 1$, let θ_n be the water level corresponding to d in the n-th order reverse waterfilling (51);
- let d_n be the distortion associated to the water level θ in the n-th order reverse waterfilling (51).

For clarity, we explicitly write down the relations between d and θ_n, and d_n and θ:

$$d = \frac{1}{n}\sum_{i=1}^{n} \min(\theta_n,\ \sigma_{n,i}^2), \tag{197}$$

$$d_n = \frac{1}{n}\sum_{i=1}^{n} \min(\theta,\ \sigma_{n,i}^2), \tag{198}$$

where $\sigma_{n,i}^2$'s are given in (60). Note that d and θ are constants independent of n, while d_n and θ_n are functions of n, and there is no direct reverse waterfilling relation between d_n and θ_n. Applying Theorem 9 in Section V-A above to the function $t \mapsto \min(\theta, \sigma^2/t)$, we have

$$\lim_{n\to\infty} d_n = d, \tag{199}$$

and

$$\lim_{n\to\infty} \theta_n = \theta. \tag{200}$$

Theorem 10 in Section V-B then implies that the convergences in (199) and (200) are both in the order of $1/n$.

A. Expectation and Variance of the d-tilted Information

Proposition 1. *For any $d \in (0, d_{\max})$ and $n \geq 1$, let d_n be defined in (198) above. Then, the expectation and variance of the d-tilted information $\jmath_{U_1^n}(U_1^n, d_n)$ at distortion level d_n satisfy*

$$\left|\frac{1}{n}\mathbb{E}\left[\jmath_{U_1^n}(U_1^n, d_n)\right] - \mathbb{R}_U(d)\right| \leq \frac{C_1}{n}, \tag{201}$$

$$\left|\frac{1}{n}\mathbb{V}\left[\jmath_{U_1^n}(U_1^n, d_n)\right] - \mathbb{V}_U(d)\right| \leq \frac{C_2}{n}, \tag{202}$$

where $\mathbb{R}_U(d)$ and $\mathbb{V}_U(d)$ are the rate-distortion function given in (53) and the informational dispersion given in (65), respectively, and C_1 and C_2 are positive constants.

Proof. Using the same derivation as that of (59), one can obtain the following representation of the d-tilted information $\jmath_{U_1^n}(U_1^n, d_n)$ at distortion level d_n:

$$\jmath_{U_1^n}(U_1^n, d_n) = \sum_{i=1}^{n} \frac{\min(\theta,\ \sigma_{n,i}^2)}{2\theta}\left(\frac{X_i^2}{\sigma_{n,i}^2} - 1\right) + \frac{1}{2}\sum_{i=1}^{n} \log\frac{\max(\theta,\ \sigma_{n,i}^2)}{\theta}, \tag{203}$$

where X_1^n is the decorrelation of U_1^n defined in (61). Note that the difference between (59) and (203) is that θ_n is replaced by θ. Using (62) and taking expectations and variances of both sides of (203), we arrive at

$$\frac{1}{n}\mathbb{E}\left[\jmath_{U_1^n}(U_1^n, d_n)\right] = \frac{1}{2n}\sum_{i=1}^{n} \log\max\left(1,\ \frac{\sigma_{n,i}^2}{\theta}\right), \tag{204}$$

$$\frac{1}{n}\mathrm{Var}\left[\jmath_{U_1^n}(U_1^n, d_n)\right] = \frac{1}{2n}\sum_{i=1}^{n} \min\left(1,\ \frac{\sigma_{n,i}^4}{\theta^2}\right). \tag{205}$$

Applying Theorem 10 in Section V-B to (204) with the function $F_{\mathrm{G}}(t)$ defined in (103) yields (201). Similarly, applying Theorem 10 to (205) with the function (174) yields (202). □

Proposition 1 is one of the key lemmas that will be used in both converse and achievability proof. Proposition 1 and its proof are similar to those of [47, Eq. (95)–(96)]. The difference is that we apply Theorem 10, which is the nonstationary version of [47, Th. 4], to a different function in (204).

B. Approximation of the d-tilted Information

The following proposition gives a probabilistic characterization on the accuracy of approximating the d-tilted information $\jmath_{U_1^n}(U_1^n, d)$ at distortion level d using the d-tilted information $\jmath_{U_1^n}(U_1^n, d_n)$ at distortion level d_n.

Proposition 2. *For any $d \in (0, d_{\max})$, there exists a constant $\tau > 0$ (depending on d only) such that for all n large enough*

$$\mathbb{P}\left[\left|\jmath_{U_1^n}(U_1^n, d) - \jmath_{U_1^n}(U_1^n, d_n)\right| > \tau\right] \leq \frac{1}{n}, \tag{206}$$

where d_n is defined in (198).

Proof. The proof in [47, App. D-B] works through for the nonstationary case as well, since the proof [47, App. D-B] only relies on that the convergences in (199) and (200) are both in the order of $1/n$, which continue to hold for the nonstationary case. □

Remark 9. The following high probability set is used in our converse and achievability proof:

$$\mathcal{A} \triangleq \left\{\left|\jmath_{U_1^n}(U_1^n, d) - \jmath_{U_1^n}(U_1^n, d_n)\right| \leq \tau\right\}. \tag{207}$$

Proposition 2 implies that $\mathbb{P}[\mathcal{A}] \geq 1 - 1/n$ for all n large enough.

APPENDIX D
CONVERSE PROOF

Proof of Theorem 6. Using the general converse by Kostina and Verdú [22, Th. 7] and our established Propositions 1 and 2 in Appendix C, the proof is the same as the converse proof in the asymptotically stationary case [47, Th. 7, Eq. (97)–(109)]. For completeness, we give a proof sketch. Choosing $\gamma = (\log n)/2$ and setting X to be U_1^n in [22, Th. 7], we know that any (n, M, d, ϵ) code for the Gauss-Markov source must satisfy

$$\epsilon \geq \mathbb{P}\left[\jmath_{U_1^n}(U_1^n, d) \geq \log M + (\log n)/2\right] - 1/\sqrt{n}. \quad (208)$$

By conditioning on the high probability set $\mathcal{A}$ defined in Remark 9 above, we can further bound ϵ from below by

$$(1 - 1/n) \cdot \mathbb{P}\left[\jmath_{U_1^n}(U_1^n, d_n) \geq \log M + (\log n)/2 + \tau\right] - 1/\sqrt{n}. \quad (209)$$

From (203), we know that $\jmath_{U_1^n}(U_1^n, d_n)$ is a sum of independent random variables with means and variances bounded by the rate-distortion function $\mathbb{R}_U(d)$ and the informational dispersion $\mathbb{V}_U(d)$, with errors in the order of $1/n$ due to Proposition 1. Choosing M as in [47, Eq. (103)] and applying the Berry-Esseen theorem to $\jmath_{U_1^n}(U_1^n, d_n)$, we obtain the converse in Theorem 6. □

APPENDIX E
ACHIEVABILITY PROOF

Proof of Theorem 7. With our lossy AEP for the nonstationary Gauss-Markov source and Propositions 1 and 2, the proof is similar to the one for the stationary Gauss-Markov source in [47, Sec. V-C]. Here, we streamline the proof. As elucidated in Section IV-E above, the standard random coding argument [22, Cor. 11] implies that for any n, there exists an (n, M, d, ϵ') code such that

$$\epsilon' \leq \inf_{P_{V_1^n}} \mathbb{E}\left[\exp\left(-M \cdot P_{V_1^n}(\mathcal{B}(U_1^n, d))\right)\right]. \quad (210)$$

Choosing V_1^n to be $V_1^{\star n}$ (the random variable that attains the minimum in (48) with X_1^n there replaced by U_1^n), the bound (210) can be relaxed to

$$\epsilon' \leq \mathbb{E}\left[\exp\left(-M \cdot P_{V_1^{\star n}}(\mathcal{B}(U_1^n, d))\right)\right]. \quad (211)$$

To simplify notations, in the following, we denote by C a constant that might be different from line to line. Given any constant $\epsilon \in (0, 1)$, define ϵ_n as

$$\epsilon_n \triangleq \epsilon - \frac{C}{\sqrt{n}} - \frac{1}{q(n)} - \frac{1}{n}, \quad (212)$$

where $q(n)$ is defined in (83) above. Note that for all n large enough, we have $\epsilon_n \in (0, 1)$. We choose M as

$$\log M \triangleq n\mathbb{R}_U(d) + \sqrt{n\mathbb{V}_U(d)}Q^{-1}(\epsilon_n) + \log(\log n/2) + p(n) + C + \tau, \quad (213)$$

where $p(n)$ is defined in (82) and τ is from Proposition 2 above. We also define the random variable G_n as

$$G_n \triangleq \log M - \jmath_{U_1^n}(U_1^n, d_n) - p(n) - C - \tau, \quad (214)$$

where d_n is define in (198) above. Note that all the randomness in G_n is from U_1^n, hence we will also use the notation $G_n(u_1^n)$ to indicate one realization of the random variable G_n. By bounding the deterministic part, that is, $\log M$, of G_n using Proposition 1, we know that with probability 1,

$$G_n \geq \mathbb{E} + Q^{-1}(\epsilon_n)\sqrt{\mathbb{V}} - \jmath_{U_1^n}(U_1^n, d_n) + \log(\log n/2), \quad (215)$$

where we use $\mathbb{E}$ and $\mathbb{V}$ to denote the expectation and variance of the informational dispersion $\jmath_{U_1^n}(U_1^n, d_n)$ at distortion level d_n. Define the set $\mathcal{G}_n$ as

$$\mathcal{G}_n \triangleq \left\{u_1^n \in \mathbb{R}^n \colon G_n(u_1^n) < \log(\log n/2)\right\}, \quad (216)$$

Then, in view of (203), the informational dispersion $\jmath_{U_1^n}(U_1^n, d_n)$ is a sum of independent random variables with bounded moments, and we apply Berry-Esseen theorem to obtain

$$P_{U_1^n}(\mathcal{G}_n) \leq \epsilon_n + \frac{C}{\sqrt{n}}. \quad (217)$$

We define one more set $\mathcal{L}_n$ as

$$\mathcal{L}_n \triangleq \left\{u_1^n \in \mathbb{R}^n \colon \log \frac{1}{P_{V_1^{\star n}}(\mathcal{B}(u_1^n, d))} < \log M - G_n(u_1^n)\right\}. \quad (218)$$

Then, by the lossy AEP in Lemma 4 in Section IV-E above and Proposition 2, we have

$$P_{U_1^n}(\mathcal{L}_n) \geq 1 - \frac{1}{q(n)} - \frac{1}{n}. \quad (219)$$

Finally, for any constant $\epsilon \in (0, 1)$ and n large enough, we define ϵ_n as in (212) above and set M as in (213). Then, there exists (n, M, d, ϵ') code such that

$$\begin{aligned}
&\epsilon' \\
&\leq \mathbb{E}\left[\exp\left(-M \cdot P_{V_1^{\star n}}(\mathcal{B}(U_1^n, d)) \cdot 1\{\mathcal{L}_n\}\right)\right] + \\
&\quad \mathbb{E}\left[\exp\left(-M \cdot P_{V_1^{\star n}}(\mathcal{B}(U_1^n, d))\right) \cdot 1\{\mathcal{L}_n^c\}\right] && (220)\\
&\leq \mathbb{E}\left[\exp(e^{-G_n})\right] + \frac{1}{q(n)} + \frac{1}{n}, && (221)
\end{aligned}$$

where the last inequality is due to the definition of $\mathcal{L}_n$ and (219). By further conditioning on $\mathcal{G}_n$, we conclude that there exists (n, M, d, ϵ') code such that

$$\begin{aligned}
\epsilon' &\leq \epsilon_n + \frac{C}{\sqrt{n}} + \frac{1}{n} + \frac{1}{q(n)} && (222)\\
&= \epsilon. && (223)
\end{aligned}$$

Therefore, by the choice of M in (213), the minimum achievable source coding rate $R(n, d, \epsilon)$ must satisfy

$$R(n, d, \epsilon) \leq \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}}Q^{-1}(\epsilon) + \frac{K_1 \log\log n}{n} + \frac{p(n)}{n} + \frac{K_2}{\sqrt{n}q(n)}, \quad (224)$$

for all n large enough, where $K_1 > 0$ is a universal constant and K_2 is a constant depending on ϵ. Here we change from $Q^{-1}(\epsilon_n)$ to $Q^{-1}(\epsilon)$ using a Taylor expansion. Therefore, Theorem 7 follows immediately from (224) with the choices of $p(n)$ and $q(n)$ given by (82) and (83), respectively, in the lossy AEP in Lemma 4 in Section IV-E above. We have $O(\cdot)$ in (78) since K_2 could be positive or negative. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

APPENDIX F
PROOF OF LOSSY AEP

A. Notations

For the optimization problem $\mathbb{R}(A_1^n, B_1^n, d)$ in (86), the generalized tilted information defined in [22, Eq. (28)] in a_1^n (a realization of A_1^n) is given by

$$\Lambda_{B_1^n}(a_1^n, \delta, d) \triangleq -\delta n d - \log \mathbb{E}\left[\exp(-n\delta \mathsf{d}\left(a_1^n, B_1^n\right))\right], \tag{225}$$

where $\delta > 0$ and $d \in (0, d_{\max})$. For properties of the generalized tilted information, see [22, App. D]. For clarity, we list the notations used throughout this section:

1) X_1^n denotes the decorrelation of U_1^n defined in (61);
2) $\hat{X}_1^n$ is the proxy random variable of X_1^n defined in Definition 2 in Section IV-F above;
3) For $Y_1^{\star n}$ that achieves $\mathbb{R}_{X_1^n}(d)$ in (48), $\hat{F}_1^{\star n}$ is the random vector that achieves $\mathbb{R}\left(\hat{X}_1^n, Y_1^{\star n}, d\right)$;
4) We denote $\lambda_n^\star$ the negative slope of $\mathbb{R}_{X_1^n}(d)$ (the same notation used in (58)):
$$\lambda_n^\star \triangleq -\mathbb{R}'_{X_1^n}(d). \tag{226}$$
It is shown in [47, Lem. 5] that $\lambda_n^\star$ is related to the n-th order water level θ_n in (51) by
$$\lambda_n^\star = \frac{1}{2\theta_n}. \tag{227}$$
Given any source outcome u_1^n, let x_1^n be the decorrelation of u_1^n. Define $\hat{\lambda}_n$ as
$$\hat{\lambda}_n \triangleq -\mathbb{R}'(\hat{X}_1^n, Y_1^{\star n}, d). \tag{228}$$
5) Comparing the definitions of d-tilted information and the generalized tilted information, one can see that [47, Eq. (18)]
$$\jmath_{X_1^n}(x_1^n, d) = \Lambda_{Y_1^{\star n}}(x_1^n, \lambda_n^\star, d). \tag{229}$$
6) Recalling (62) and applying the reverse waterfilling result [67, Th. 10.3.3], we know that the coordinates of $Y_1^{\star n}$ are independent and satisfy
$$Y_i^\star \sim \mathcal{N}(0,\ \nu_{n,i}^2), \tag{230}$$
where
$$\nu_{n,i}^2 \triangleq \max(0,\ \sigma_{n,i}^2 - \theta_n), \tag{231}$$
with $\theta_n > 0$ given in (197).

B. Parametric Representation of the Gaussian Conditional Relative Entropy Minimization

Various aspects of the optimization problem (86) have been discussed in [47, Sec. II-B]. In particular, let $B_1^{\star n}$ be the optimizer of $\mathbb{R}_{A_1^n}(d)$, then we have

$$\mathbb{R}(A_1^n, B_1^{\star n}, d) = \mathbb{R}_{A_1^n}(d), \tag{232}$$

where $\mathbb{R}_{A_1^n}(d)$ is in (48). Another useful result on the optimization problem (86) is the following: when the input A_1^n and B_1^n are independent Gaussian random vectors, we have parametric characterizations for the optimizer and optimal value of (86), using the negative slope of $\mathbb{R}(A_1^n, B_1^n, d)$ w.r.t. d. This result is summarized in the following theorem.

Theorem 11. *Let $A_1, \ldots, A_n$ be independent random variables with*

$$A_i \sim \mathcal{N}(0, \alpha_i^2), \tag{233}$$

and $B_1, \ldots, B_n$ be independent random variables with

$$B_i \sim \mathcal{N}(0, \beta_i^2),. \tag{234}$$

For any d such that

$$0 < d < \frac{1}{n}\sum_{i=1}^n (\alpha_i^2 + \beta_i^2), \tag{235}$$

we have the following parametric representation for $\mathbb{R}(A_1^n, B_1^n, d)$:

$$\mathbb{R}(A_1^n, B_1^n, d) = -\lambda d + \frac{1}{2n}\sum_{i=1}^n \log(1 + 2\lambda\beta_i^2) + \frac{1}{n}\sum_{i=1}^n \frac{\lambda\alpha_i^2}{1 + 2\lambda\beta_i^2} \tag{236}$$

$$d = \frac{1}{n}\sum_{i=1}^n \frac{\alpha_i^2 + \beta_i^2(1 + 2\lambda\beta_i^2)}{(1 + 2\lambda\beta_i^2)^2}, \tag{237}$$

where $\lambda > 0$ is the parameter.

Similar results to Theorem 11 have appeared previously in the literature [20, 37, 42]. See [37, Example 1 and Th. 2] for the case of $n = 1$. For completeness, we present a proof.

Proof. Fix any d that satisfies (235), and let λ be such that (237) is satisfied. Note from (237) that d is a strictly decreasing function in λ (unless $\beta_i = 0$ for all $i \in [n]$), hence such λ is unique. The upper bound on d in (235) guarantees that $\lambda > 0$. We first show the $\leq$ direction in (236). For $A_1^n = a_1^n \in \mathbb{R}^n$, define the conditional distribution $P_{F_i|A_i}(f_i|a_i)$ as

$$\mathcal{N}\left(\frac{2\lambda\beta_i^2 a_i}{1 + 2\lambda\beta_i^2}, \frac{\beta_i^2}{1 + 2\lambda\beta_i^2}\right). \tag{238}$$

We then define the joint distribution $P_{A_1^n, F_1^n}$ as

$$P_{A_1^n, F_1^n} \triangleq \prod_{i=1}^n P_{F_i|A_i} P_{A_i}. \tag{239}$$

Using (237), we can check that with such a choice of $P_{A_1^n, F_1^n}$, the expected distortion between A_1^n and F_1^n equals d. The details follow.

$$\begin{aligned}
&\mathbb{E}[\mathsf{d}\left(A_1^n, F_1^n\right)] \\
=&\mathbb{E}\left\{\mathbb{E}[\mathsf{d}\left(A_1^n, F_1^n\right)|A_1^n]\right\} && (240)\\
=&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\mathbb{E}[(F_i - A_i)^2|A_i]\right\} && (241)\\
=&\frac{1}{n}\sum_{i=1}^n \frac{\beta_i^2}{1 + 2\lambda\beta_i^2} + \frac{\alpha_i^2}{(1 + 2\lambda\beta_i^2)^2} && (242)\\
=&d, && (243)
\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

where (242) is from the relation $\mathbb{E}[(X-t)^2] = \mathrm{Var}[X] + (\mathbb{E}[X]-t)^2$ and (243) is due to (237). Therefore, the choice of $P_{F_1^n|A_1^n}$ in (238) and (239) is feasible for the optimization problem in defining $\mathbb{R}(A_1^n, B_1^n, d)$. Hence,

$$\mathbb{R}(A_1^n, B_1^n, d) \leq \frac{1}{n} D(P_{F_1^n|A_1^n} || P_{B_1^n} | P_{A_1^n}) \tag{244}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[D(P_{F_i|A_i}(\cdot|A_i) || P_{B_i})]. \tag{245}$$

It is straightforward to verify that the Kullback-Leibler divergence between two Gaussian distributions $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ is given by

$$D(P_X||P_Y) = \frac{\sigma_X^2 + (\mu_X - \mu_Y)^2}{2\sigma_Y^2} - \frac{1}{2}\log\frac{\sigma_X^2}{\sigma_Y^2} - \frac{1}{2}. \tag{246}$$

Using (246) and (238), we see that (245) equals the right-hand side of (236). To prove the other direction, we use the Donsker-Varadhan representation of the Kullback-Leibler divergence [68, Th. 3.5]:

$$D(P||Q) = \sup_g \ \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[\exp g(X)], \tag{247}$$

where the supremum is over all functions g from the sample space to $\mathbb{R}$ such that both expectations in (247) are finite. Fix any $P_{F_1^n|A_1^n}$ such that $\mathbb{E}[\mathsf{d}(A_1^n, F_1^n)] \leq d$. For any $A_1^n = a_1^n$, in (247), we choose P to be $P_{F_1^n|A_1^n}(\cdot|a_1^n)$, Q to be $P_{B_1^n}$ and g to be $g(f_1^n) \triangleq -n\lambda \mathsf{d}(f_1^n, a_1^n)$ for any $f_1^n \in \mathbb{R}^n$, then we have

$$D(P_{F_1^n|A_1^n}(\cdot|a_1^n)||P_{B_1^n}) \geq -n\lambda \mathbb{E}_{P_{F_1^n|A_1^n}(\cdot|a_1^n)}[\mathsf{d}(F_1^n, a_1^n)] - \log \mathbb{E}_{P_{B_1^n}}[\exp\{-n\lambda \mathsf{d}(B_1^n, a_1^n)\}]. \tag{248}$$

Taking expectations on both sides of (248) with respect to $P_{A_1^n}$ and then normalizing by n, we have

$$\mathbb{R}(A_1^n, B_1^n, d) \geq -\lambda \mathbb{E}[\mathsf{d}(A_1^n, F_1^n)] - \mathbb{E}_{P_{A_1^n}} \log \mathbb{E}_{P_{B_1^n}}[\exp\{-n\lambda \mathsf{d}(B_1^n, A_1^n)\}]. \tag{249}$$

Using the formula for the moment generating function for noncentral χ^2 distributions, we can compute

$$\mathbb{E}_{P_{B_1^n}}[\exp\{-n\lambda \mathsf{d}(B_1^n, a_1^n)\}] = \prod_{i=1}^{n} \frac{1}{\sqrt{1+2\lambda\beta_i^2}} \exp\left(\frac{-\lambda a_i^2}{1+2\lambda\beta_i^2}\right). \tag{250}$$

Plugging (250) into (249) and using $\mathbb{E}[\mathsf{d}(A_1^n, F_1^n)] \leq d$, we conclude that $\mathbb{R}(A_1^n, B_1^n, d)$ is greater than or equal to the right-hand side of (236). □

Our next result states that for fixed β_i^2's satisfying certain mild conditions, if we change the variances from α_i^2's to $\hat{\alpha}_i^2$'s, then the perturbation on the corresponding λ's is controlled by the perturbation on α_i^2's.

Theorem 12 (Variance perturbation)**.** *Let α_i^2's and β_i^2's be in (233) and (234) above, respectively. For a fixed d satisfying (235), let λ be given by (237). Suppose that α_i^2's and β_i^2's are such that both*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{(1+2\lambda\beta_i^2)^4} \tag{251}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \frac{2\beta_i^2(2\alpha_i^2 + 1 + 2\lambda\beta_i^2)}{(1+2\lambda\beta_i^2)^3} \tag{252}$$

are bounded by positive constants. Let $\hat{A}_1, \ldots, \hat{A}_n$ be independent random variables with

$$\hat{A}_i \sim \mathcal{N}(0, \hat{\alpha}_i^2). \tag{253}$$

Let $\hat{\lambda}$ be such that

$$d = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\alpha}_i^2 + \beta_i^2(1+2\hat{\lambda}\beta_i^2)}{(1+2\hat{\lambda}\beta_i^2)^2}. \tag{254}$$

Then, there is a constant $C > 0$ such that

$$\left|\hat{\lambda} - \lambda\right| \leq C \max_{1\leq i\leq n} \left|\hat{\alpha}_i^2 - \alpha_i^2\right|. \tag{255}$$

Proof. We can view (237) as an equation of the form $f(\alpha_1^2, \ldots, \alpha_n^2, \lambda) = 0$. Then, by the implicit function theorem, we know that there exists a unique continuously differentiable function h such that

$$\lambda = h(\alpha_1^2, \ldots, \alpha_n^2), \tag{256}$$

and

$$\frac{\partial h}{\partial \alpha_i^2} = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{2\beta_i^2[2\alpha_i^2 + \beta_i^2(1+2\lambda\beta_i^2)]}{(1+2\lambda\beta_i^2)^3}\right\}^{-1} \frac{1}{n(1+2\lambda\beta_i^2)^2}. \tag{257}$$

Hence,

$$\|\nabla h\|_2 = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{2\beta_i^2(2\alpha_i^2 + 1 + 2\lambda\beta_i^2)}{(1+2\lambda\beta_i^2)^3}\right\}^{-1} \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} \frac{1}{(1+2\lambda\beta_i^2)^4}}. \tag{258}$$

By the assumptions (251) and (252), we know that there exists a constant $C > 0$ such that

$$\|\nabla h\|_2 \leq \frac{C}{\sqrt{n}}. \tag{259}$$

Hence, we have

$$\left|\hat{\lambda} - \lambda\right| \leq \|\nabla h\|_2 \|(\alpha_1^2, \ldots, \alpha_n^2) - (\hat{\alpha}_1^2, \ldots, \hat{\alpha}_n^2)\|_2 \tag{260}$$

$$\leq C \max_{1\leq i\leq n} \left|\hat{\alpha}_i^2 - \alpha_i^2\right|. \tag{261}$$

□

C. Proof of Theorem 8

The proof is similar to [47, Th. 12], and we streamline the proof and point out the differences. We use the notations defined in Appendix F-A above.

Our Corollary 1 implies that for all n large enough the condition (89) is violated with probability at most $2e^{-cn}$ for a constant $c > \log(a)/2$. This is much stronger than the bound $\Theta\left(1/\text{poly}\log n\right)$ in the stationary case [47, Th. 6].

In view of (62), the random variables $X_i/\sigma_{n,i}$ for $i = 1, \ldots, n$, are distributed according to i.i.d. standard normal distributions and their $2k$-th moments equal to $(2k-1)!!$. The Berry-Esseen theorem implies that the condition (90) is violated with probability at most $\Theta\left(1/\sqrt{n}\right)$. This is the same as the stationary case [47, Eq. (279)–(280)].

We use the following procedure to show that the condition (91) is violated with probability at most $\Theta\left(1/\log n\right)$:

- We approximate $m_i(u_1^n)$ by another random variable $\bar{m}_i(u_1^n)$ that is easier to analyze.
- We show that (91) with $m_i(u_1^n)$ replaced by $\bar{m}_i(u_1^n)$ holds with probability at least $1 - \Theta(1/\log n)$.
- We then control the difference between $m_i(u_1^n)$ and $\bar{m}_i(u_1^n)$.

To carry out the above program, we first give an expression for $m_i(u_1^n)$ by applying [47, Lem. 4] (see also the proof of Theorem 11) on $\mathbb{R}(\hat{X}_1^n, Y_1^{\star n}, d)$. Note that $\hat{X}_1^n$ and $Y_1^{\star n}$ are Gaussian random vectors with independent coordinates with variances given by (85) and (230), respectively. Then, [47, Lem. 4] implies that the optimizer $P_{\hat{F}_1^{\star n}|\hat{X}_1^n}$ for $\mathbb{R}(\hat{X}_1^n, Y_1^{\star n}, d)$ satisfies

$$P_{\hat{F}_1^{\star n}|\hat{X}_1^n=\hat{x}_1^n} = \prod_{i=1}^n P_{\hat{F}_i^{\star}|\hat{X}_i=\hat{x}_i}, \tag{262}$$

where the conditional distributions $\hat{F}_i^{\star}|\hat{X}_i = \hat{x}_i$ are Gaussian:

$$\mathcal{N}\left(\frac{2\hat{\lambda}_n\nu_{n,i}^2\hat{x}_i}{1+2\hat{\lambda}_n\nu_{n,i}^2}, \frac{\nu_{n,i}^2}{1+2\hat{\lambda}_n\nu_{n,i}^2}\right), \tag{263}$$

where $\nu_{n,i}^2$'s are defined in (231) above and $\hat{\lambda}_n$ is defined in (228) above. Then, using the definition of $m_i(u_1^n)$ in (87) and (263) above, we obtain

$$m_i(u_1^n) = \frac{\nu_{n,i}^2}{1+2\hat{\lambda}_n\nu_{n,i}^2} + \frac{x_i^2}{(1+2\hat{\lambda}_n\nu_{n,i}^2)^2}, \tag{264}$$

where $x_1^n = \mathsf{S}^\top u_1^n$. The random variable $m_i(u_1^n)$ in the form of (264) is hard to analyze since we do not have a simple expression for $\hat{\lambda}_n$. By replacing $\hat{\lambda}_n$ with $\lambda_n^\star$, we define another random variable $\bar{m}_i(u_1^n)$ that turns out to be easier to analyze:

$$\bar{m}_i(u_1^n) \triangleq \frac{\nu_{n,i}^2}{1+2\lambda_n^\star\nu_{n,i}^2} + \frac{x_i^2}{(1+2\lambda_n^\star\nu_{n,i}^2)^2}. \tag{265}$$

Plugging (227) and (231) into (265), we obtain

$$\bar{m}_i(u_1^n) = \frac{\min(\sigma_{n,i}^2, \theta_n)^2}{\sigma_{n,i}^2}\left(\frac{x_i^2}{\sigma_{n,i}^2} - 1\right) + \min(\sigma_{n,i}^2, \theta_n), \tag{266}$$

where θ_n is the n-th order water level in (51) and $x_1^n = \mathsf{S}^\top u_1^n$. The random variable $\bar{m}_i(U_1^n)$ is much easier to analyze since $X_i/\sigma_{n,i}$'s are i.i.d. standard normal random variables. Moreover, in view of (51), their expectations satisfy

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\bar{m}_i(U_1^n)\right] = \frac{1}{n}\sum_{i=1}^n \min(\sigma_{n,i}^2, \theta_n) = d. \tag{267}$$

Since $X_i/\sigma_{n,i}$'s have bounded moments, from Berry-Esseen theorem, we know that there exists a constant $\omega > 0$ such that for all n large enough

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n \bar{m}_i(U_1^n) - d\right| > \omega\eta_n\right] \le \frac{C_1}{\log n} + \frac{C_2}{\sqrt{n}}, \tag{268}$$

where η_n is in (88) above and C_1, C_2 are positive constants. In the last step of the program, we control the difference between $m_i(U_1^n)$ and $\bar{m}_i(U_1^n)$. From (264)–(265), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \bar{m}_i(u_1^n) - \frac{1}{n}\sum_{i=1}^n m_i(u_1^n) \\
=&\frac{1}{n}\sum_{i=1}^n \frac{2\nu_{n,i}^4(\hat{\lambda}_n - \lambda_n^\star)}{(1+2\hat{\lambda}_n\nu_{n,i}^2)(1+2\lambda_n^\star\nu_{n,i}^2)} + \\
&\frac{1}{n}\sum_{i=1}^n \frac{2x_i^2\nu_{n,i}^2(2+2\hat{\lambda}_n\nu_{n,i}^2+2\lambda_n^\star\nu_{n,i}^2)(\hat{\lambda}_n - \lambda_n^\star)}{(1+2\hat{\lambda}_n\nu_{n,i}^2)^2(1+2\lambda_n^\star\nu_{n,i}^2)^2}.
\end{aligned} \tag{269}$$

For $i = 1$, we have $\nu_{n,1}^2 = \sigma_{n,1}^2 - \theta_n = \Theta\left(a^{2n}\right)$, $\hat{\lambda}_n = \Theta(1)$ and $\lambda_n^\star = \Theta(1)$. This implies that the summands in (269) for $i = 1$ are both of the order $O(1/n)$ for any $x_1^2 = O(a^{4n})$. For $2 \le i \le n$, the condition (89) and the variance perturbation result in Theorem 12 imply that every summand in (269) for $i \ge 2$ is in the order of η_n. Hence, (269) is in the order of η_n. Finally, combining (268) and (269) implies that conditioning on the conditions (89) and (90), we have (91) is violated with probability at most $\Theta(1/\log n)$.

□

D. Auxiliary Lemmas

Lemma 7 (Lower bound on the probability of distortion balls)**.** *Fix $d \in (0, d_{\max})$. For any n large enough and any $u_1^n \in \mathcal{T}(n,p)$ defined in Definition 3 in Section IV-F above, and γ defined in (298) below, it holds that*

$$\mathbb{P}\left[d - \gamma \le \mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right) \le d \,|\hat{X}_1^n = x_1^n\right] \ge \frac{K_1}{\sqrt{n}}, \tag{270}$$

where $K_1 > 0$ is a constant and $\hat{F}_1^{\star n}$ is in Appendix F-A above.

The proof is in Appendix F-F.

Lemma 8. *Fix $d \in (0, d_{\max})$ and $\epsilon \in (0,1)$. There exists constants C and $K_2 > 0$ such that for all n large enough,*

$$\begin{aligned}
&\mathbb{P}\left[\Lambda_{Y_1^{\star n}}\left(X_1^n, \hat{\lambda}_n, d\right) \le \Lambda_{Y_1^{\star n}}\left(X_1^n, \lambda_n^\star, d\right) + C\log n\right] \\
\ge& 1 - \frac{K_2}{\sqrt{n}},
\end{aligned} \tag{271}$$

where $\lambda_n^\star$ and $\hat{\lambda}_n$ are defined in (226) and (228), respectively.

Proof. The proof of Lemma 8 is the same as [47, Eq. (314)–(333)] except that we strengthen the right side of [47, Eq. (322)]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2021.3050342, IEEE Transactions on Information Theory

to be $\Theta(e^{-cn})$ for a constant $c > \log(a)/2$ due to Corollary 1. □

E. Proof of Lemma 4

Using Lemmas 7 and 8 in Appendix F-D above, the proof of Lemma 4 is almost the same as that in the stationary case [47, Eq. (270)-(278)]. For completeness, we sketch the proof here. We weaken the bound [22, Lem. 1] by setting $P_{\hat{X}}$ as $P_{\hat{X}_1^n}$ and P_Y as $P_{Y_1^{\star n}}$ to obtain that for any $x_1^n \in \mathbb{R}^n$,

$$\begin{aligned}&\log \frac{1}{P_{Y_1^{\star n}}(\mathcal{B}_d(x_1^n))}\\ &\le \inf_{\gamma>0} \ \Lambda_{Y_1^{\star n}}(x_1^n, \hat{\lambda}_n, d) + \hat{\lambda}_n n\gamma - \\ &\qquad \log \mathbb{P}\left[d-\gamma \le \mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right) \le d | \hat{X}_1^n = x_1^n\right], \end{aligned} \tag{272}$$

where $\hat{\lambda}_n$ in (228) depends on X_1^n. Let $\mathcal{E}$ denote the event inside the square brackets in (81). Then,

$$\begin{aligned}&\mathbb{P}[\mathcal{E}]\\ =&\mathbb{P}[\mathcal{E} \cap \mathcal{T}(n,p)] + \mathbb{P}[\mathcal{E} \cap \mathcal{T}(n,p)^c] &(273)\\ \le&\mathbb{P}\Big[\Lambda_{Y_1^{\star n}}(X_1^n, \hat{\lambda}_n, d) \ge \Lambda_{Y_1^{\star n}}(X_1^n, \lambda_n^\star, d) + p(n) - \hat{\lambda}_n n\gamma - \\ &\qquad \frac{1}{2}\log n + \log K_1, \ \mathcal{T}(n,p)\Big] + \mathbb{P}[\mathcal{T}(n,p)^c] &(274)\\ \le&\mathbb{P}\Big[\Lambda_{Y_1^{\star n}}(X_1^n, \hat{\lambda}_n, d) \ge \Lambda_{Y_1^{\star n}}(X_1^n, \lambda_n^\star, d) + C\log n\Big] + \\ &\qquad \mathbb{P}[\mathcal{T}(n,p)^c] &(275)\\ \le&\frac{1}{q(n)}, &(276)\end{aligned}$$

where

- (274) is due to (272) and Lemma 7;
- From (274) to (275), we used the fact that for $u_1^n \in \mathcal{T}(n,p)$, $\hat{\lambda}_n$ can be bounded by

$$\left|\hat{\lambda}_n - \frac{1}{2\theta}\right| \le B_1, \tag{277}$$

  where $B_1 > 0$ is a constant and $\theta > 0$ is given by (54). The bound (277) is obtained by the same argument as that in the stationary case [47, Eq. (273)]; γ is chosen in (298) below; the constants c_i's, $i = 1, ...4$ in (82) are chosen as

$$c_1 = B_1 + \frac{1}{2\theta}, \tag{278}$$
$$c_2 = B_4, \tag{279}$$
$$c_3 = C + \frac{1}{2}, \tag{280}$$
$$c_4 = -\log K_1, \tag{281}$$

  where $B_4 > 0$ is given in (297) below and K_1 and C are the same constants in Lemmas 7 and 8, respectively.
- (276) is due to Lemma 8 and Theorem 8.

□

F. Proof of Lemma 7

Proof. The proof is similar to the stationary case [47, Lem. 10]. We streamline the proof and point out the differences. Conditioned on $\hat{X}_1^n = x_1^n$, the random variable

$$\mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right) = \frac{1}{n}\sum_{i=1}^n \left(\hat{F}_i^\star - x_i\right)^2 \tag{282}$$

follows a noncentral χ^2-distribution with (at most) n degrees of freedom, since it is shown in [47, Eq. (282) and Lem. 4] that conditioned on $\hat{X}_1^n = x_1^n$, the distribution of the random variable $\hat{F}_i^\star - x_i$ is given by

$$\mathcal{N}\left(\frac{-x_i}{1+2\hat{\lambda}_n\nu_i^2}, \ \frac{\nu_i^2}{1+2\hat{\lambda}_n\nu_i^2}\right), \tag{283}$$

where ν_i^2's are given in (231). Then, the conditional expectation is given by

$$\mathbb{E}\left[\mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right) | \hat{X}_1^n = x_1^n\right] = \frac{1}{n}\sum_{i=1}^n m_i(u_1^n), \tag{284}$$

where $m_i(u_1^n)$ is defined in (87) in Section IV-E above. In view of (282), (284) and (91), we expect that $\mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right)$ concentrates around d conditioned on $\hat{X}_1^n = x_1^n$ for $u_1^n \in \mathcal{T}(n,p)$. Note that the proof of Theorem 8 related to (91) is different from the one in the stationary case, see Appendix F-C above for the details. To simplify notations, we denote the variances as

$$V_i(x_1^n) \triangleq \mathrm{Var}\left[\left(\hat{F}_i^\star - x_i\right)^2 | \hat{X}_1^n = x_1^n\right], \tag{285}$$

$$V(x_1^n) \triangleq \sqrt{\frac{1}{n}\sum_{i=1}^n V_i(x_1^n)}. \tag{286}$$

Due to (283) and (91), we see $(\hat{F}_i^\star - x_i)^2$'s have finite second- and third- order absolute moments. That is, we have

$$V(x_1^n) = \Theta(1), \tag{287}$$

for $u_1^n \in \mathcal{T}(n,p)$. Therefore, we can apply the Berry-Esseen theorem. Hence,

$$\begin{aligned}&\mathbb{P}\left[d-\gamma \le \mathsf{d}\left(x_1^n, \hat{F}_1^{\star n}\right) \le d \ | \hat{X}_1^n = x_1^n\right]\\ =&\mathbb{P}\Bigg[\frac{n(d-\gamma) - \sum_{i=1}^n m_i(u_1^n)}{\sqrt{n}V(x_1^n)}\\ &\qquad \le \frac{1}{\sqrt{n}V(x_1^n)}\sum_{i=1}^n\left[\left(\hat{F}_i^\star - x_i\right)^2 - m_i(u_1^n)\right]\\ &\qquad \le \frac{nd - \sum_{i=1}^n m_i(u_1^n)}{\sqrt{n}V(x_1^n)} \ | \hat{X}_1^n = x_1^n\Bigg] &(288)\\ \ge&\Phi\left(\frac{nd - \sum_{i=1}^n m_i(u_1^n)}{\sqrt{n}V(x_1^n)}\right)\\ &\qquad - \Phi\left(\frac{n(d-\gamma) - \sum_{i=1}^n m_i(u_1^n)}{\sqrt{n}V(x_1^n)}\right) - \frac{2B_1}{\sqrt{n}} &(289)\\ =&\frac{\sqrt{n}\gamma}{V(x_1^n)}\Phi'(\xi) - \frac{2B_1}{\sqrt{n}}, &(290)\end{aligned}$$

where

- (289) follows from the Berry-Esseen theorem; $B_1 > 0$ is a constant, and

$$\Phi(t) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{\tau^2}{2}} \, d\tau \tag{291}$$

  is the cumulative distribution function of the standard Gaussian distribution;
- (290) is due to the mean value theorem and

$$\Phi'(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}; \tag{292}$$

- In (290), ξ satisfies

$$\frac{n(d-\gamma) - \sum_{i=1}^{n} m_i(u_1^n)}{\sqrt{n} V(x_1^n)} \leq \xi \leq \frac{nd - \sum_{i=1}^{n} m_i(u_1^n)}{\sqrt{n} V(x_1^n)}. \tag{293}$$

By (91) and (287), we see that there is a constant $B_2 > 0$ such that

$$\left| \frac{nd - \sum_{i=1}^{n} m_i(u_1^n)}{\sqrt{n} V(x_1^n)} \right| \leq B_2 \sqrt{\log \log n}. \tag{294}$$

Hence, as long as γ in (293) satisfies

$$\gamma \leq O(\eta_n), \tag{295}$$

where η_n is defined in (88), there exists a constant $B_3 > 0$ such that

$$|\xi| \leq B_3 \sqrt{\log \log n}. \tag{296}$$

Let $B_4 > 0$ be a constant such that

$$B_4 \geq \frac{B_3^2}{2} + 1, \tag{297}$$

and choose γ as

$$\gamma \triangleq \frac{(\log n)^{B_4}}{n}, \tag{298}$$

which satisfies (295). Then, plugging the bounds (287), (296), (297) and (298) into (290), we conclude that there exists a constant $K_1 > 0$ such that (290) is further bounded from below by $\frac{K_1}{\sqrt{n}}$. □

REFERENCES

[1] P. Tian and V. Kostina, "From parameter estimation to dispersion of nonstationary Gauss-Markov processes," in *Proceedings of 2019 IEEE International Symposium on Information Theory*, Paris, France, Jul. 2019, pp. 2044–2048.

[2] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica, Journal of the Econometric Society*, pp. 173–220, Jul. 1943.

[3] H. Rubin, "Consistency of maximum likelihood estimates in the explosive case," *Statistical Inference in Dynamic Economic Models*, pp. 356–364, 1950.

[4] J. S. White, "The limiting distribution of the serial correlation coefficient in the explosive case," *The Annals of Mathematical Statistics*, pp. 1188–1197, Dec. 1958.

[5] T. W. Anderson, "On asymptotic distributions of estimates of parameters of stochastic difference equations," *The Annals of Mathematical Statistics*, pp. 676–687, Sep. 1959.

[6] J. Rissanen and P. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes," *The Annals of Statistics*, pp. 297–315, Mar. 1979.

[7] N. H. Chan and C.-Z. Wei, "Asymptotic inference for nearly nonstationary AR(1) processes," *The Annals of Statistics*, pp. 1050–1063, Sep. 1987.

[8] B. Bercu, F. Gamboa, and A. Rouault, "Large deviations for quadratic forms of stationary Gaussian processes," *Stochastic Processes and their Applications*, vol. 71, no. 1, pp. 75–90, Oct. 1997.

[9] J. Worms, "Large and moderate deviations upper bounds for the Gaussian autoregressive process," *Statistics & probability letters*, vol. 51, no. 3, pp. 235–243, Feb. 2001.

[10] A. Rantzer, "Concentration bounds for single parameter adaptive control," in *Proceedings of 2018 IEEE Annual American Control Conference*, Milwaukee, WI, USA, Jun. 2018, pp. 1862–1866.

[11] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 1, pp. 142–163, Mar. 1959.

[12] T. Goblick, "A coding theorem for time-discrete analog data sources," *IEEE Transactions on Information Theory*, vol. 15, no. 3, pp. 401–407, May 1969.

[13] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Transactions on Information Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.

[14] T. Berger, "Information rates of Wiener processes," *IEEE Transactions on Information Theory*, vol. 16, no. 2, pp. 134–139, Mar. 1970.

[15] R. M. Gray, "Information rates of autoregressive processes," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 412–421, Jul. 1970.

[16] T. Hashimoto and S. Arimoto, "On the rate-distortion function for the nonstationary Gaussian autoregressive process," *IEEE Transactions on Information Theory*, vol. 26, no. 4, pp. 478–480, Jul. 1980.

[17] R. M. Gray and T. Hashimoto, "A note on rate-distortion functions for nonstationary Gaussian autoregressive processes," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1319–1322, Feb. 2008.

[18] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 197–199, Mar. 1974.

[19] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion. I. known statistics," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 71–91, Jan. 1997.

[20] E.-H. Yang and Z. Zhang, "On the redundancy of lossy source coding with abstract alphabets," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1092–1110, May 1999.

[21] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Proceedings of 2011 IEEE Data Compression Conference*, Snowbird, UT, USA, Mar. 2011, pp. 53–62.

[22] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012.

[23] T. Haavelmo, "The statistical implications of a system of simultaneous equations," *Econometrica, Journal of the Econometric Society*, pp. 1–12, Jan. 1943.

[24] T. Koopmans, "Serial correlation and quadratic forms in normal variables," *The Annals of Mathematical Statistics*, vol. 13, no. 1, pp. 14–33, Mar. 1942.

[25] J. P. Gould and C. R. Nelson, "The stochastic structure of the velocity of money," *The American Economic Review*, vol. 64, no. 3, pp. 405–418, Jun. 1974.

[26] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 427–431, Jun. 1979.

[27] P. Whittle, *Hypothesis testing in time series analysis*. Almqvist & Wiksells boktr., 1951, vol. 4.

[28] G. E. P. Box and G. M. Jenkins, *Time series analysis: forecasting and control*. San Francisco: Holden-Day, 1970.

[29] A. H. S. Kailath, Thomas and B. Hassibi, *Linear estimation*. New Jersey: Prentice Hall, 2000.

[30] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, New Jersey: P T R Prentice Hall, 1987.

[31] S. L. Tu, "Sample complexity bounds for the linear quadratic regulator," Ph.D. dissertation, UC Berkeley, 2019.

[32] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, Apr. 1971.

[33] T. Berger, "Rate distortion theory for sources with abstract alphabets and memory," *Information and Control*, vol. 13, no. 3, pp. 254–273, Sep. 1968.

[34] R. M. Gray, "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Transactions on Information Theory*, vol. 17, no. 2, pp. 127 – 134, Mar. 1971.

[35] A. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Transactions on Information Theory*, vol. 17, no. 5, pp. 508–513, Sep. 1971.

[36] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, Jan. 2000.

[37] A. Dembo and L. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.

[38] I. Kontoyiannis, "Pattern matching and lossy data compression on random fields," *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 1047–1051, Apr. 2003.

[39] I. Kontoyiannis and R. Zamir, "Mismatched codebooks and the role of entropy coding in lossy data compression," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1922–1938, May. 2006.

[40] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: rate-distortion theorems and error exponents for a general source," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2154–2179, Jun. 2007.

[41] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory," *IEEE Transactions on information theory*, vol. 48, no. 8, pp. 2276–2290, Aug. 2002.

[42] A. Dembo and I. Kontoyiannis, "The asymptotics of waiting times between stationary processes, allowing distortion," *Annals of Applied Probability*, pp. 413–429, May. 1999.

[43] M. Madiman, M. Harrison, and I. Kontoyiannis, "Minimum description length vs. maximum likelihood in lossy data compression," in *Proceedings of 2004 IEEE International Symposium on Information Theory*, Chicago, IL, USA, Jun. 2004, p. 461.

[44] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.

[45] V. Y. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 881–903, Feb. 2014.

[46] S. Watanabe, "Second-order region for Gray–Wyner network," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 1006–1018, Feb. 2017.

[47] P. Tian and V. Kostina, "The dispersion of the Gauss-Markov source," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6355–6384, Oct. 2019.

[48] L. Zhou, V. Y. Tan, and M. Motani, "Discrete lossy Gray-Wyner revisited: Second-order asymptotics, large and moderate deviations," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1766–1791, Mar. 2016.

[49] ——, "Second-order and moderate deviations asymptotics for successive refinement," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2896–2921, Mar. 2017.

[50] S. Tatikonda, A. Sahai, and S. Mitter, "Stochastic linear control over a communication channel," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1549–1561, Sep. 2004.

[51] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, Nov. 2019.

[52] R. M. Gray, "In memory of A.H. "Steen" Gray Jr." *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 96–100, Mar. 2020.

[53] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 06–09 Jul. 2018, pp. 439–473.

[54] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *2019 American Control Conference (ACC)*, Philadelphia, USA, Jul. 2019, pp. 5655–5661.

[55] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun. 2019, pp. 5610–5618.

[56] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, Oct. 2018.

[57] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Berlin: Springer-Verlag, 2010.

[58] B. Bercu and A. Touati, "Exponential inequalities for self-normalized martingales with applications," *The Annals of Applied Probability*, vol. 18, no. 5, pp. 1848–1869, 2008.

[59] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, New Jersey: Prentice Hall, 1971.

[60] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2019, vol. 48.

[61] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[62] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints–II: Stabilization with limited information feedback," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 1049–1053, May 1999.

[63] J. Baillieul, "Feedback designs for controlling device arrays with communication channel bandwidth constraints," in *Proceedings of 1999 ARO Workshop on Smart Structures*, Pennsylvania State University, State College, PA, USA, Aug. 1999, pp. 48–55.

[64] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, Jul. 2004.

[65] R. M. Gray, "Toeplitz and Circulant Matrices: A Review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[66] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*. New York: Chelsea Publishing Company, 1984.

[67] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Nov. 2012.

[68] Y. Polyanskiy and Y. Wu, *Lecture notes on information theory*. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf

Peida Tian is a Ph.D candidate in the Department of Electrical Engineering at California Institute of Technology. He received a B. Engg. in Information Engineering and a B. Sc. in Mathematics from the Chinese University of Hong Kong (2016), and a M.S. in Electrical Engineering from Caltech (2017). He is interested in optimization and information theory.

Victoria Kostina (S'12–M'14) received the bachelor's degree from the Moscow Institute of Physics and Technology in 2004, the master's degree from the University of Ottawa in 2006, and the Ph.D. degree from Princeton University in 2013. She was affiliated with the Institute for Information Transmission Problems, Russian Academy of Sciences. In 2014, she joined Caltech, where she is currently a Professor of electrical engineering. Her research spans information theory, coding, control, learning, and communications. She received the Natural Sciences and Engineering Research Council of Canada master's scholarship in 2009, the Princeton Electrical Engineering Best Dissertation Award in 2013, the Simons-Berkeley Research Fellowship in 2015, and the NSF CAREER Award in 2017.